86-02959/

Environment Canada
Environnement Canada

A Program for Estimating the Point of Change and Degree in Polynomial Regression

S.R. Esterby

SCIENTIFIC SERIES NO. 147

INLAND WATERS DIRECTORATE
NATIONAL WATER RESEARCH INSTITUTE
CANADA CENTRE FOR INLAND WATERS
BURLINGTON, ONTARIO, 1985

(Disponible en français sur demande)

© Minister of Supply and Services Canada 1985

Cat. No. En36-502/147E

ISBN 0-662-14377-9

Contents

	Page
ABSTRACT	v
RÉSUMÉ	v
INTRODUCTION	1
ALGORITHM	2
EXAMPLES	4
Example 1	4
Example 2	8
Example 3	10
Example 4	10
DISCUSSION	12
ACKNOWLEDGMENTS	12
REFERENCES	12
APPENDIX	13

Tables

1. User-specified program options	2
2. Differences owing to the assumption about the variances	4
3. Values of program options specified for the examples	5

Illustrations

Figure 1. Flow chart showing the main features of the program	3
Figure 2. Printer output for first run of core KB, Example 1	6
Figure 3. Plots for core KB, Example 1	8
Figure 4. Plot of the data with fitted curves and the relative likelihood function, Example 2	9
Figure 5. Plots for the normal data, Example 3	10
Figure 6. Plots constructed from runs 2 and 3 of Example 4	11

Abstract

A Fortran computer program, which implements an estimation procedure for a polynomial regression model with a change point, is described. The assumptions of the model are analogous to those for a single polynomial but include the assumption that the polynomial, and possibly the variance, have changed at a point in the data set. The order of the data points must also be known. The degree of the polynomial in each segment and the point of change are estimated by an iterative procedure which involves the use of the marginal likelihood function for the change-point parameter and a test using the change in residual sums of squares with reduction of degree. The assumptions of the model can be examined using the residuals and the likelihood function. Examples are given to demonstrate the utility of the program and to show the output from both the printer and the plotter.

Résumé

On décrit un programme informatique en Fortran qui utilise une méthode de calcul pour un modèle de régression polynomiale avec un point de transition. Les hypothèses posées pour le modèle sont les mêmes que pour un seul polynôme, mais on suppose que le polynôme, et peut-être la variance, changent en un point de la série de données. L'ordre des données simples doit aussi être connu. Le degré du polynôme dans chaque segment et le point de transition sont calculés par une méthode itérative qui comporte l'emploi de la fonction de vraisemblance marginale pour le paramètre du point de transition et un test faisant appel au changement des sommes des carrés des écarts résiduels avec une diminution du degré. Les hypothèses du modèle peuvent être examinées à l'aide des valeurs des écarts résiduels et de la fonction de vraisemblance. On donne des exemples qui démontrent l'utilité du programme et indiquent les résultats obtenus avec l'imprimante et le traceur.

A Program for Estimating the Point of Change and Degree in Polynomial Regression

S.R. Esterby

INTRODUCTION

Regression analysis is often used to provide information about the relationship between variables. In most applications, one regression relationship is fitted to a set of data, but situations arise where it is known that the relationship has changed during the period of data collection. The complexity of the analysis depends upon the degree of knowledge about the form of the relationship and the point of change. If the point at which the regression relationship changes is known, then the analysis consists of fitting regular least-squares relationships to the separate segments, possibly subject to the constraint that the regression functions agree at the point of change. If the point of change is unknown, then the analysis is more difficult, since the change point must be estimated as well as the regression parameters. The present report addresses this latter situation and describes an algorithm based on a procedure given by Esterby and El-Shaarawi (1981a).

The algorithm has been written for regression on one dependent variable. The usual assumptions of least-squares regression are made within each segment of the data fitted by one regression function, plus the assumption that the regression function has changed at an unknown point. The form of the regression function is restricted to polynomials, and an ordering of the observations, such as a natural ordering in time or space, is required. Let (x_i, y_i) be independent pairs of observations, where the index i denotes the order of observation. Then the model consists of polynomials of degree p and q, with the change from a polynomial of degree p to a polynomial of degree q occurring after the observation with $i=n_1$. This is given by

$$y_i = \sum_{j=0}^{p} \theta_{1j} x_i^j + e_{1i} \qquad i = 1,2...n_1$$

and

$$y_i = \sum_{j=0}^{q} \theta_{2j} x_i^j + e_{2i} \qquad i = n_1+1, n_1+2...n$$

where $e_{1i}(i=1,2...n_1)$ and e_{2i} $(i=n_1+1, n_1+2...n)$ are two independent sequences of random variables which are assumed to be $N(0, \sigma_1^2)$ and $N(0, \sigma_2^2)$, respectively. The parameters $\underline{\theta}_1 = (\theta_{10}, \theta_{11}...\theta_{1p})$, $\underline{\theta}_2 = (\theta_{20}, \theta_{21}...\theta_{2q})$, σ_1^2, σ_2^2 and n_1 are unknown. In the next two sections, this model is denoted by $M_u(p,q|n_1)$. Given p and q, the change-point parameter, n_1, is estimated from the marginal likelihood function of n_1. If p and q are also unknown, an iterative procedure is used to estimate p and q. The procedure combines the marginal likelihood function of n_1 and a test analogous to that for reduction of degree in a single polynomial. A maximum value for the degree of each polynomial is assumed. Let (p^*,q^*) be the pair of maximum values, then the total degree $(p + q)$ is reduced by 1 at each step in the iteration, until no further reduction is possible or $(p,q) = (0,0)$. The case in which $\sigma_1^2 = \sigma_2^2$ is also considered, and this model is denoted by $M_e(p,q|n_1)$, where the subscripts e and u are used to distinguish models obtained under the different variance assumptions. There are differences between the marginal likelihood functions and the procedures for estimating p and q in the cases where the variances are assumed equal and unequal. Details of the methods are given in Esterby and El-Shaarawi (1981a), and a more detailed account of likelihood functions can be found in Esterby and El-Shaarawi (1981b).

The methods of residual analysis, which are available for checking the assumptions of the model when a single function is fitted, can be used here by combining the residuals from each segment. In addition, examination of the likelihood function permits a check on the assumption of one point of change.

The major objective of this procedure is to estimate the point at which the regression relationship changes when it can be assumed that one change has occurred. No constraint on the values of the regression functions at the change point is imposed. Other objectives may require different analyses. If a smooth curve fitted to the data is required, fitting either a higher order polynomial or a cubic spline to the data would be appropriate in many cases.

A description of the algorithm is given in the next section, and examples of the use of the program in the following section. The Fortran code has not been listed because the program is very long. A copy of the program and test cases can be obtained by writing to the author. The program has been written in Fortran V, using some IMSL subroutines, and implemented on a CDC Cyber 171 with the NOS operating system. Plotting programs were written for a CALCOMP 1036 plotter. The purpose here is to describe the features of the program without reiterating the development of the procedure, and to illustrate its use by means of examples, so that the reader in need of such a program can evaluate its usefulness for a particular application.

ALGORITHM

Many features have been included in the program to enable the user to analyze simply and efficiently one or more pairs of variables in one computer job, provided that all the variables are on the same file. The options that must be specified by the user are listed in Table 1. The program can estimate the point of change, n_1, given the degrees of the polynomials, or estimate the degrees of the polynomials and the point of change, given the maximum value which the degree can assume in each segment. For each pair of variables, the user can specify either the single pair of degrees, (p,q), or up to five values for the maximum degrees (p^*,q^*). For the single pair (p,q) or for each pair (p^*,q^*), the user must specify the assumption about the equality of variances, whether plots are to be produced and the two critical significance levels for the tests of hypotheses. Different plots are produced when the independent variable is not in ascending or descending order.

Figure 1 gives the main features of the program and shows where the user-specified options determine which parts of the program are used. To understand the details of the calculations performed for each step shown in the flow chart, it is necessary to read the paper by Esterby and El-Shaarawi (1981a). Several comments will help clarify the logic. The iterative procedure for estimation of p and q first assumes particular values for p and q and estimates n_1 as $\hat{n}_1$. Then holding n_1 fixed at $\hat{n}_1$, a test is performed to determine whether the total degree (p+q) can be reduced by 1. If the iteration continues, that is, if the degree was found to be (p+q-1), then n_1 is again estimated assuming the new pair of degrees. The two sequences shown in Figure 1, for the cases of equal and unequal variances, are further explained in Table 2, where the differences owing to the two cases are listed. The estimate of n_1, $\hat{n}_1$, is the value of n_1 which maximizes the marginal likelihood function and this is determined by calculating the marginal likelihood function for n_1 equal to the index values shown in Table 2.

Several of the steps shown in Figure 1 were not included in the procedure given in the 1981a paper by Esterby and El-Shaarawi. To test for equality of variances (see A in Fig. 1), it is assumed that the variances are unequal, and the point of change is estimated using the marginal likelihood function for unequal variances. A standard F-test for equality of variances is then applied to test the hypothesis that the residual variances from segments 1 and 2 are equal. During the estimation of p and q, if the degree in one segment reaches 0, then it is necessary to test for reduction in the other segment only. The test becomes identical with testing for reduction of degree when a single polynomial is fit to a set of data. Thus the critical significance level α is used directly, whereas in the case of two possible ways in which the degree can be reduced by 1, $\sqrt{\alpha}$ is used.

Implementation of the algorithm also required several additional steps. To permit matrix inversion, the independent variable is standardized. Further calculations are required to obtain the regression parameters in terms of the original independent variable from the parameters in terms

Table 1. User-Specified Program Options

Level of application	Options
	Number of pairs of variables
For each variable pair	Format for reading variable values, x then y
	Title
	Whether the independent variable is ordered
	Either (1) the known values of (p,q) or
	(2) the number of pairs of maximum degree for the polynomials, n_{pq}, and their values (p^*,q^*)
	Whether all the samples are to be used or which ones are to be omitted
	Units for the axes of the plots
For the single pair (p,q) or for each pair of (p^*,q^*)	Assumption about the variances (equal, unequal or test for equality)
	Whether plots are to be produced
	Critical significance level for the test for equality of variances
	Critical significance level for the test for reduction of degree

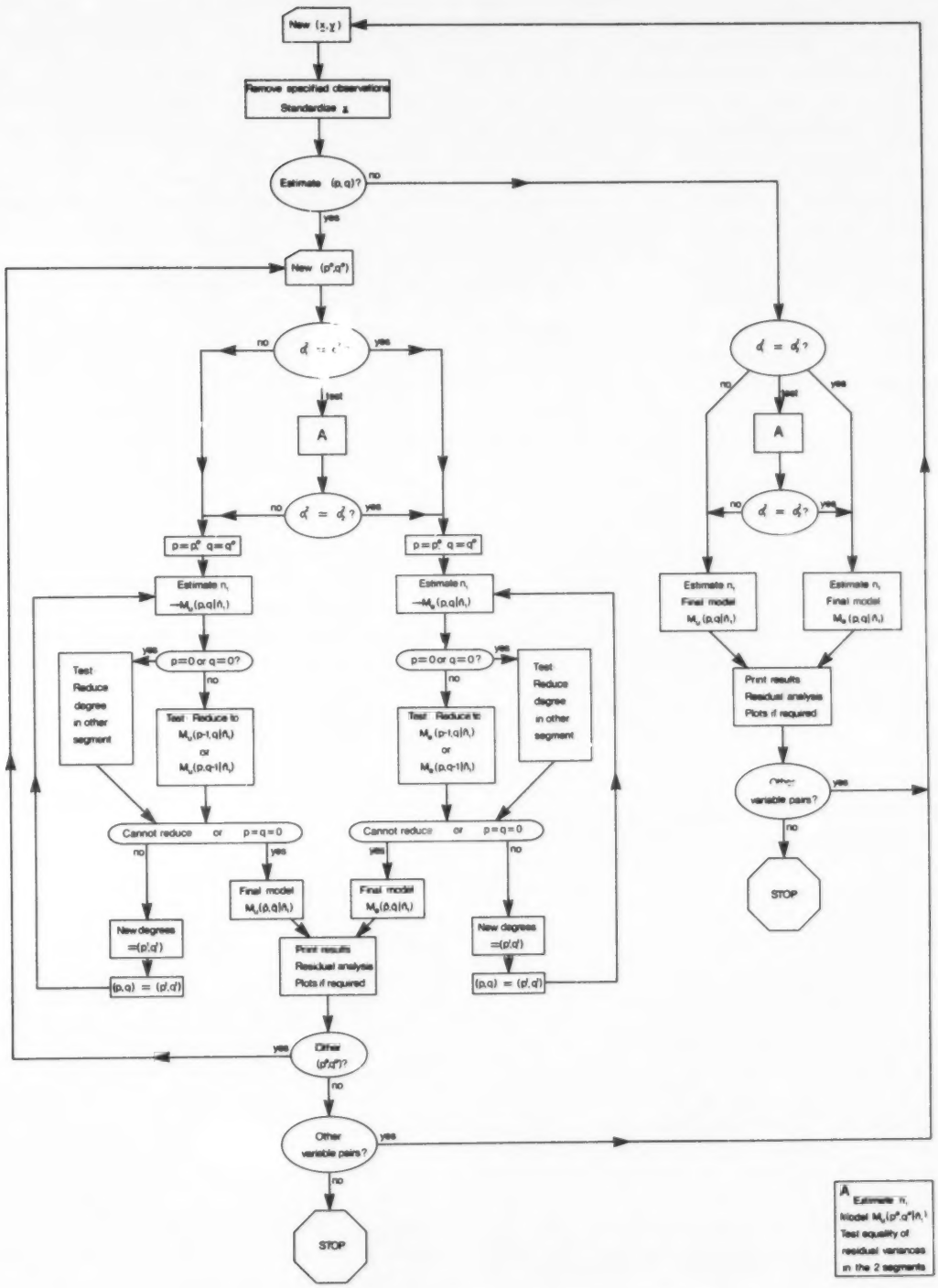


Figure 1. Flow chart showing the main features of the program.

Table 2. Differences Owing to the Assumption about the Variances

Feature	Equal variances	Unequal variances
Permissible values of n_1	p+1. . .n-(q+1)	p+2. . .n-(q+2)
Marginal likelihood function of n_1, $L(n_1)$	$\{\nu_1\hat{\sigma}_1^2+\nu_2\hat{\sigma}_2^2\}^{-1/2(\nu_1+\nu_2-1)}$	$\frac{\Gamma(1/2\nu_1)\ \Gamma(1/2\nu_2)}{\nu_1^{1/2\nu_1}\ \nu_2^{1/2\nu_2}}\ \hat{\sigma}_1^{-(\nu_1-1)}\ \hat{\sigma}_2^{-(\nu_2-1)}$
Test for reduction of degree, neither p=0 nor q=0	Residual sums of squares from the two segments are pooled to estimate the common variance	Residual sum of squares for each segment is used separately to estimate the different variances
Standardized residuals	Residuals from both segments are divided by $\hat{\sigma}$ obtained from the pooled residual sums of squares	Residuals are standardized first by dividing by $\hat{\sigma}_1$ or $\hat{\sigma}_2$ and then combined

Note: $\nu_1 = n_1-(p+1)$, $\nu_2 = (n-n_1) - (q+1)$, $\hat{\sigma}_1^2 = \sum_{i=1}^{n_1} (y_i - \sum_{j=0}^{p} \hat{\theta}_{1j}\, x_i^j)^2/\nu_1$

$\hat{\sigma}_2^2 = \sum_{i=n_1+1}^{n} (y_i - \sum_{j=0}^{q} \hat{\theta}_{2j}\, x_i^j)^2/\nu_2$ and $\Gamma(z)$ is the gamma function of z.

of the standardized variable. The formulae are given in the Appendix. In the case of equal variances, when neither p nor q is 0, the test for reduction of degree uses the sum of the residual sums of squares in the denominator of the test statistic. Thus the lower and upper limits for n_1 can be p+1 and n-(q+1), resulting in the degrees of freedom for the residual sum of squares in one segment being equal to 0. However, if this occurs when the degree is 0 in the other segment, where the test is based on the segment with non-zero degree only, a test statistic with denominator degrees of freedom equal to 0 results. The correction of this situation, which has been implemented, is to adjust the corresponding limit for n_1 so that there is always one observation more than the number of parameters to be estimated, and to maintain this adjustment for the rest of the iteration. For equal variances and the number of parameters equal to the number of observations in one segment, the residual sum of squares is not calculated but assigned its theoretical value of 0. A zero residual sum of squares can also occur if all the values of the dependent variable are equal in one segment. Assigning to the residual sum of squares the theoretical value of 0 in the case of unequal variances is not satisfactory, since the logarithm of the residual sum of squares for each segment is needed for the marginal likelihood function. For this case, the residual sum of squares in the segment with constant dependent variable is assigned the arbitrary value of 0.000 000 1. This has the effect of treating constancy in the same manner for equal and unequal variances. The user can circumvent this arbitrary decision simply by specifying that the constant end values be omitted.

The results which are printed include a summary of the steps in the estimation of p and q, and for the final model, the relative marginal likelihood function of n_1, the regression parameters, $\underline{\hat{\theta}}_1$ and $\underline{\hat{\theta}}_2$, the predicted values of $\underline{y}$, $\underline{\hat{y}}$, and the residuals $\underline{e}=\underline{y}-\underline{\hat{y}}$. The residual sums of squares, residual mean squares, total corrected sum of squares and R^2, as defined in the Appendix, are also printed. The residual analysis (Draper and Smith, 1981) is largely graphical, consisting of a Q-Q plot to check the assumption of normality, and plots of the standardized residuals versus $\underline{x}$ and $\underline{\hat{y}}$. A test for runs in the residuals is also performed.

EXAMPLES

Examples were used to assess the performance of the procedure in previous publications (Esterby and El-Shaarawi, 1980, 1981a,b). For the examples given here, the use of the computer program, including its model-checking capabilities, is emphasized. Example 1 is discussed in detail and the program's printer and plot output are shown. In the other three examples, the plots are limited to those which illustrate particular features of the example.

Example 1

The rise of *Ambrosia* pollen towards the surface of a sediment core can be used as a time marker when the corresponding date of forest clearing is known. The assumptions of one change and of a polynomial functional form for the regression relationship between concentration and depth are tenable. The data set used here was given by Robbins, Edgington and Kemp (1978) and consists of *Ambrosia* concentrations (grains·g^{-1} of dry sediment) and sediment depth (cm) for 23 samples of a sediment core

from Lake Ontario. The midpoint of the sample interval was used as sample depth. Sample order was assigned from top to bottom, with the surface sample as sample one. The values of the options used to run the computer program are given in Table 3, and the resulting computer listing and plots in Figures 2 and 3.

Table 3. Values of Program Options Specified for the Examples

Option	Examples						
	1a	1b	2	3	4a	4b	4c
x Ordered	Yes	Yes	No	Yes	Yes	Yes	Yes
(p,q) or (p*,q*)	(1,1)	(1,0)	(1,1)	(0,0)	(3,3)	(1,1)	(0,1)
Samples used	All	All	All	All	All	1-44	17-49
Variances	Test	Test	Test	Equal	Test	Test	Test
α, Variances	0.05	0.05	0.05	0.05	0.05	0.05	0.05
α, Degree	0.01	0.01	0.01	0.01	0.01	0.01	0.01
Plots	Yes	Yes	Yes	Yes	No	Yes	Yes
x Origin	0.0	0.0	0.0	0.0		0.0	0.0
limit	68.4	68.4	18.0	100.0		800.0	800.0
increment	7.6	7.6	2.25	10.0		100.0	100.0
y Origin	0.0	0.0	0.0	-3.3		0.0	0.0
limit	47.5	47.5	20.04	3.3		20.4	20.4
increment	9.5	9.5	3.34	1.1		3.4	3.4

Note: The letters a,b,c refer to runs within the example, thus 1a is the first program run described in Example 1.

The printer output has extensive titles and should be easy to understand (Fig. 2). The conclusion from the test for equality of variances applies to all successive results. The relative likelihood function, residual sums of squares and mean squares, fitted model and residuals are printed for the final model only. The last part of the output is a summary of the final model. The factor listed there must be applied to all residual sums of squares and mean squares in the previous results.

The final model was $M_u(1,1|15)$, that is, two straight lines, with different coefficients for samples 1 to 15 and samples 16 to 23, and unequal variances in the two segments. Since pollen analysts generally describe the changes from the bottom to the top of the core, this model is one of an increase in concentration at the rate of 0.10 grains·g^{-1}·cm^{-1} from the bottom sample to sample 16, and, at a higher rate, 0.94 grains·g^{-1}·cm^{-1} for samples 15 to the surface. Higher variation was also exhibited in the upper 15 samples, with $\hat{\sigma}_1^2$=41.29 and $\hat{\sigma}_2^2$=3.20. The rates and variances are given as the second parameter based on untransformed x and the residual mean square for each segment in Figure 2. The point of change, sample 15, is the sample for which the relative likelihood function attains its maximum value of one. The strength of the evidence that sample 15 is the change point, relative to all other samples permitted by the model, is obtained by examining the relative likelihood function, $R(n_1)$. In this case all samples in the closed interval [12,17] have a relative likelihood function $\geq$0.10. By quoting, for example, the 0.10 relative likelihood interval, a measure of the uncertainty of the estimate is given.

For the independent variable, in ascending or descending order, four plots are produced. Figure 3a contains the plot of the data with the fitted curves superimposed and the relative likelihood function. The left vertical and lower horizontal axes are for the data and fitted curves, and the right vertical and upper horizontal axes for the relative likelihood function for the change-point parameter, n_1. Note that only integer values of n_1 can have non-zero relative likelihood, but the points for the relative likelihood function have been joined for ease of separation of the two types of information in the plot. The residual plots, shown in Figures 3b, c, and d, indicate no marked departures from the assumptions of the model. The standardized residual of largest magnitude is that of sample 6 at -2.5, and under normality one expects that about 98% of the residuals will be in the interval [-2.5, 2.5]. Furthermore, the observed significance level of 0.97 for the runs test (Fig. 2) provides no evidence against the model assumption of independence.

The broad likelihood interval for the point of change would be of concern in using the *Ambrosia* horizon to date this core. As discussed by Robbins *et al.*, a clear definition of what is meant by the *Ambrosia* rise or horizon has not been given. If it is defined as the point at which the concentration rises above background levels, then this means that q=0 in the present method. When (p,q) was specified as (1,0), the model which resulted (Fig. 3e) was $M_u(1,0|17)$ with a 0.10 relative likelihood interval of [15, 18]. There were no marked differences in the residual analysis and $100R^2$=86.0 and 86.8 for $M_u(1,1|15)$ and $M_u(1,0|17)$, respectively. In Robbins *et al.*, the *Ambrosia* rise had been set at sample 15 by inspection of the data. This corresponds to the first model estimated here, that of a change from one linear increase to another, with higher rate. Comparison of results from the present procedure with the poorly defined visual method should help the analyst establish a better defined procedure for determining *Ambrosia* horizons.

In general, the specification of (p*,q*) or the decision to assume (p,q) known depends upon the particular situation. In the *Ambrosia* example, to assume higher values for

S.E.G.R.E.G

DATA SET 1

FORMAT FOR READING X AND Y: (TR6,F6.2,TL11,F5.1)

CORE KB AMBROSIA

X IS IN EITHER INCREASING OR DECREASING ORDER

P AND Q WILL BE ESTIMATED

23 PAIRS OF (X,Y) VALUES READ:

.500	46.600
1.500	31.900
3.000	37.500
5.000	30.600
7.000	34.400
9.000	12.900
11.000	30.300
13.000	24.300
15.000	20.400
17.000	17.400
19.000	27.000
21.000	24.000
23.000	12.900
25.000	14.430
27.000	11.400
29.000	6.000
35.000	6.400
42.500	2.100
47.500	2.600
52.500	1.300
57.500	2.000
62.500	4.900
67.500	1.200

DATA IS NOT MANIPULATED

MEAN OF X VALUES = 25.69565

STANDARD DEVIATION OF X VALUES = 20.36359

0 ADJACENT EQUAL VALUES AT BEGINNING OF Y VECTOR

0 ADJACENT EQUAL VALUES AT END OF Y VECTOR

PLOT SPECIFICATIONS

XORIGIN= 0.000 XLIMIT= 68.400 XSTEP= 7.630 YORIGIN= 0.000 YLIMIT= 47.500 YSTEP= 9.500

PAIR 1

MAXIMUM P = 1 MAXIMUM Q = 1

EQUALITY OF VARIANCES TO BE TESTED

PLOTS WILL BE PRODUCED

SIGNIFICANCE LEVEL FOR EQUALITY OF VARIANCES TEST = .05000

SIGNIFICANCE LEVEL FOR DETERMINATION OF DEGREES OF POLYNOMIALS = .01000

TEST FOR EQUALITY OF VARIANCES

SEGMENT	DEGREES OF FREEDOM	RESIDUAL SS	RESIDUAL MS
1	13	5.3674925157	.4128840397
2	6	.1920299883	.0320048480

F RATIO = 12.90067

SIGNIFICANCE LEVEL = .00246486

VARIANCES ARE UNEQUAL

ESTIMATION OF P AND Q, UNEQUAL VARIANCES

N1HAT	MODEL 1 P	MODEL 1 Q	MODEL 2 P	MODEL 2 Q	INCREASE IN RSS	F RATIO	DF	OBS S.L.	MAX S.L.	THEOR S.L.	CONCLUDE
15	1	1	0	1	9.496601	23.00065	1,13	.00035			
			1	0	.122858	3.83874	1, 5	.09779	.09779	.10000	ADEQUATE

THE THEORETICAL SIGNIGICANCE LEVEL IS THE SQUARE ROOT OF SPECIFIED SIGNIGICANCE LEVEL IF THERE ARE TWO POSSIBLE WAYS TO REDUCE THE DEGREE; OTHERWISE IT IS THE SPECIFIED SIGNIFICANCE LEVEL.

TABLE OF RESIDUAL SUMS OF SQUARES AND MEAN SQUARES CALCULATED DURING THE ESTIMATION OF P AND Q

N1HAT	P	Q	RSS1	RSS2	RMS1	RMS2
15	1	1	5.3674925157	.1920299883	.4128840397	.0320048480
15	0	0	14.8640933333	.3148875000	1.0617209524	.0449839286

Figure 2. Printer output for first run of core KB, Example 1.

P = 1 Q = 1

N1	RESIDUAL SUMS OF SQUARES SEGMENT 1	RESIDUAL SUMS OF SQUARES SEGMENT 2	RESIDUAL MEAN SQUARES SEGMENT 1	RESIDUAL MEAN SQUARES SEGMENT 2	LIKELIHOOD FUNCTION (RELATIVE)
3	.0047605264	6.4240926888	.0047605264	.3568940338	.0135031626
4	.8095717392	6.2467126821	.4047858696	.3674536990	.0066384656
5	1.1529138989	5.1183713469	.3843046330	.3198982092	.0159189212
6	2.1584752322	4.1290101579	.5396188081	.2752673439	.0220677984
7	3.4917569930	3.4508413604	.6983513986	.2464886686	.0177281441
8	3.6681152344	3.2788882589	.6113525391	.2522221738	.0126664386
9	3.6960207362	3.2537019916	.5280029623	.2711418243	.0077657073
10	3.7031005032	3.2512143972	.4628875629	.2955649452	.0055898927
11	4.8302442976	2.0412903486	.5366938106	.2041290949	.0129163538
12	5.2489541575	.7822390625	.5248954158	.0869154514	.2676507197
13	5.3612703328	.7255364874	.4873882121	.0907046109	.1376431183
14	5.3634444412	.4193465694	.4469537034	.0539066528	.3402528632
15	5.3674925157	.1920293803	.4128840397	.0320048480	1.0000000000
16	5.4944660498	.1835927918	.3924619178	.0367185584	.3119150932
17	5.5603852142	.0903771429	.3712256809	.0225942857	.3174742161
18	5.7461140819	.0903600000	.3591321301	.0301200000	.0052934910
19	6.2395189234	.0871200000	.3670305249	.0435600000	.0210392022
20	6.7739165209	.3726000000	.3763298067	.0726000000	.0086527419

N1HAT	RESS1	RESS2	RMS1	RMS2
15	5.3674925157	.1920293803	.4128840397	.0320048480

I	Y=F(THETA)	Y=F(BETA)
1	36.94	36.94
W	1.50	1.50

ORDER OF OBSERVATION	OBSERVED X	OBSERVED Y	ESTIMATED Y	RESIDUAL	ORDERED STANDARDIZED RESIDUAL	STANDARD NORMAL QUANTILE
1	.5000000000	46.6000000000	36.9382360283	9.6617639717	-2.4937503539	-2.0190862006
2	1.5000000000	31.3000000000	35.3953677411	-4.0953677411	-1.0519682562	-1.6123892061
3	3.0000000000	37.5000000000	34.5810553103	2.9189346897	-.9447605875	-1.2334945620
4	5.0000000000	30.0000000000	32.6953287359	-2.6953287359	-.6373512728	-1.0271542729
5	7.0000000000	34.4000000000	30.8095921615	3.5904078385	-.6195304061	-.8572543202
6	9.0000000000	12.9000000000	28.9238555871	-16.0238555871	-.4952872339	-.7091025035
7	11.0000000000	30.3000000000	27.0381190127	3.2618809873	-.4461285360	-.5751092013
8	13.0000000000	24.3000000000	25.1523824383	-.8523824383	-.4394449110	-.4507442126
9	15.0000000000	20.4000000000	23.2666458639	-2.8666458639	-.4194660975	-.3330057982
10	17.0000000000	17.4000000000	21.3809092895	-3.9809092895	-.2762846346	-.2197229150
11	19.0000000000	27.0000000000	19.4951727151	7.5048272849	-.1690769659	-.1092004817
12	21.0000000000	24.3000000000	17.6094361407	6.3905638593	-.1326540292	0.0000000000
13	23.0000000000	12.9000000000	15.7236995663	-2.8236995663	-.0859415399	.1092004817
14	25.0000000000	14.4000000000	13.8379629919	.5620370081	.3974683362	.2197229150
15	27.0000000000	11.4000000000	11.9522264174	-.5522264174	.3796095133	.3330057982
16	29.0000000000	6.0000000000	5.3208817054	.6791182946	.4542661019	.4507442126
17	35.0000000000	6.4000000000	4.7258056445	1.6741943555	.5076374291	.5751092013
18	42.5000000000	2.1000000000	3.9819605685	-1.8819605685	.5547656959	.7091025035
19	47.5000000000	2.6000000000	3.4860638511	-.8860638511	.9358322093	.8572543202
20	52.5000000000	1.3000000000	2.9901671337	-1.6901671337	.9945465871	1.0271542729
21	57.5000000000	2.0000000000	2.4942704163	-.4942704163	1.1679564632	1.2334945620
22	62.5000000000	4.9000000000	1.9983736990	2.9016263010	1.5036348270	1.6123892061
23	67.5000000000	1.2000000000	1.5024769816	-.3024769816	1.6219355555	2.0190862006

SUM OF RESIDUALS, SEGMENT 1 .0000000000

SUM OF RESIDUALS, SEGMENT 2 .0000000000

TEST FOR RUNS IN RESIDUALS

NUMBER OF RUNS	16
NUMBER OF POSITIVE SIGNS	10
NUMBER OF NEGATIVE SIGNS	13
MEAN	12.3
STANDARD DEVIATION	2.30
N(0,1) VARIATE	1.82
PROBABILITY LEVEL	.97

FINAL MODEL

ALL RESIDUAL SUMS OF SQUARES AND RESIDUAL MEAN SQUARES WERE DIVIDED BY 100.

POINT OF CHANGE PARAMETER = 15

DEGREE OF POLYNOMIAL, SEGMENT 1 = 1

DEGREE OF POLYNOMIAL, SEGMENT 2 = 1

REGRESSION PARAMETERS

A. TRANSFORMED X						
SEGMENT 1	13.18205	-19.20019	0.00000	0.00000	0.00000	0.00000
SEGMENT 2	5.64866	-2.01965	0.00000	0.00000	[illegible]	0.00000
B. UNTRANSFORMED X						
SEGMENT 1	37.40967	-.94287	0.00000	0.00000	0.00000	0.00000
SEGMENT 2	8.19708	-.09918	0.00000	0.00000	[illegible]	0.00000

100 R SQUARED = 86.02406

TOTAL SUM OF SQUARES = 3977.92435

Figure 2. Continued.

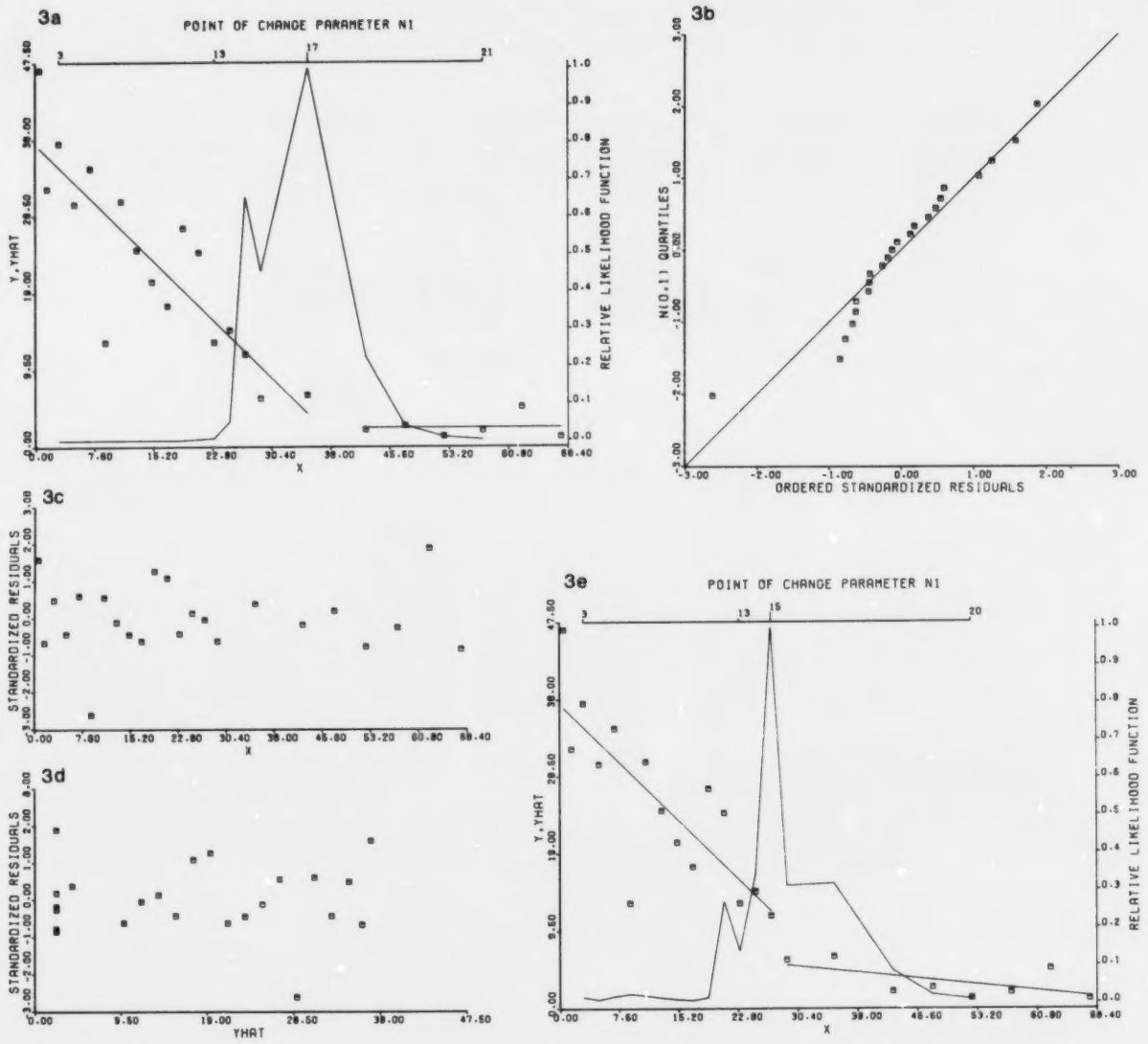


Figure 3. Plots for core KB, Example 1.

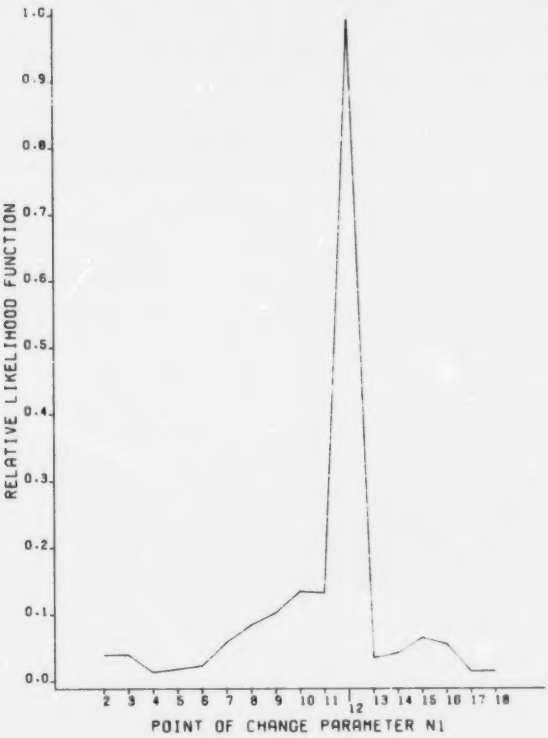
(p^*,q^*) would have resulted in fitting smaller excursions in the data and failed to address the question of a rise due to increased local input of *Ambrosia* pollen to the lake sediments. However, from the author's experience in analyzing other cores, it is clear that no one assumption would be appropriate for all cores.

Example 2

In the previous example, the order of the independent variable and the order of observation were the same and thus the two regression models were based on non-overlapping ranges for the independent variable. Although this is more commonly encountered, situations arise where the ranges of the independent variable overlap. Such a situation could occur, for example, in the study of the stability of the relationship between two variables over time, where the value of the independent variable is not controlled, and at some point a different person started taking the measurements.

The artificial data generated by Quandt (1958) were analyzed using the options specified in Table 3. The purpose here is to show the two plots produced for the data

and fitted models and for the relative likelihood function (Fig. 4). Different symbols are used for the data points and fitted lines in the two segments and the relative likelihood function is plotted in the same manner as in the case where only one plot is produced. The model corresponding to the plot is $M_e(1,1|12)$, with $\hat{y} = 2.22 + 0.69x$ and $\hat{y} = 5.91 + 0.48x$ being the fitted models for samples 1 to 12 and 13 to 20, respectively.



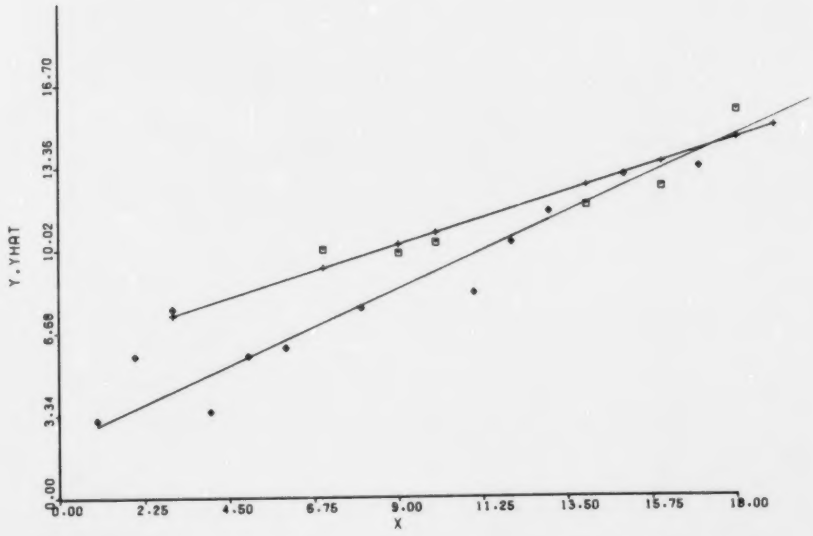


Figure 4. Plot of the data with fitted curves and the relative likelihood function, Example 2.

Example 3

The procedure does not include a test for the existence of a change. It assumes that change occurs. To show what might be expected when the procedure is applied to data in which no change has occurred, the program was run on a sample of 100 random variables generated from a N(0,1) distribution. The program options are listed in Table 3 and the plots are shown in Figure 5. Since all possible values for the change-point parameter have a value of the relative likelihood function ≥ 0.30, clearly the correct conclusion, of no point of change, would be drawn.

Example 4

Although the method is designed for estimating only one point of change, several ad hoc procedures are possible when more than one change occurs. One of the ad hoc procedures is shown here using the percentages of *Pinus* pollen in a sediment core from Bog D Pond, Hubbard County, Minnesota. The percentages were calculated from the counts reported by McAndrews (1966). The percentage of *Pinus banksiana/resinosa* type declined sharply from high values in the lowest samples, whereas *Pinus strobus* increased towards the surface of the core. The sum of these

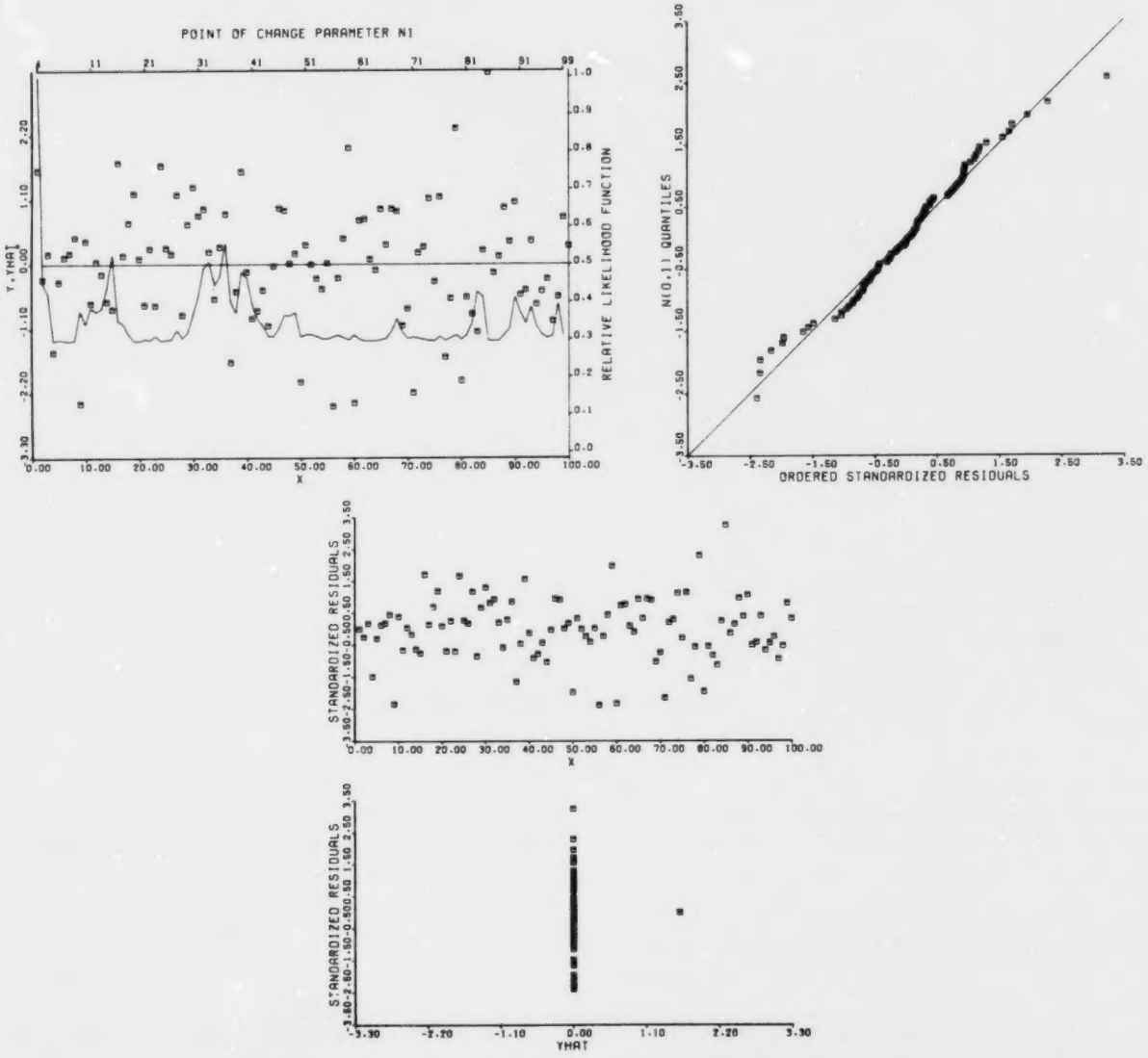


Figure 5. Plots for the normal data, Example 3.

two percentages should thus exhibit two changes. The program was run three times to reach a final model for the sum of the two percentages. The steps were as follows:

(1) All 49 samples were used and the model $M_e(2,1|44)$ was obtained.

(2) From samples 1 to 44, the model was $M_u(0,0|16)$.

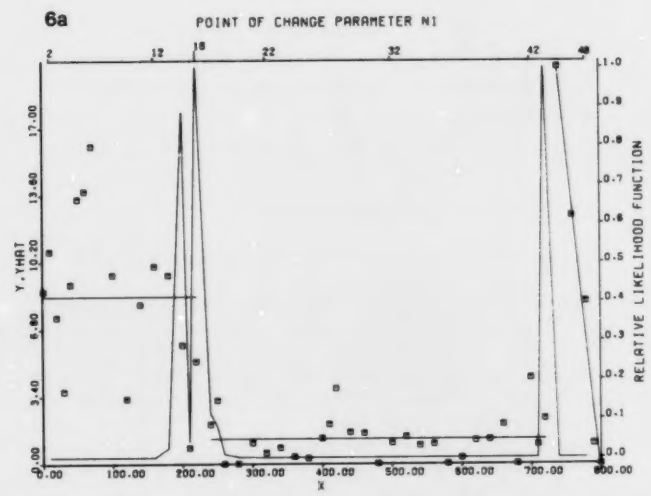
(3) To ascertain that the lower point of change is 44, a sequence was analyzed containing only one change, namely samples 17 to 49. The result was the model $M_u(0,1|44)$.

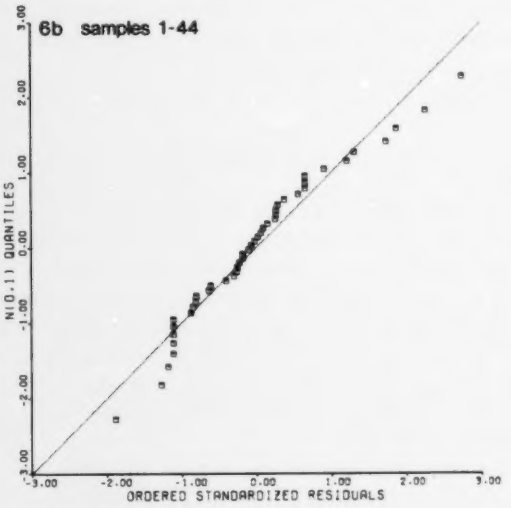
The program options are given in Table 3. Since the same scales were specified in the plots for the last two runs, composite plots have been constructed by cutting and pasting together the regions 1 to 16 and 17 to 49 as shown in three of the plots (Figs. 6a,c,d). The separate Q-Q plots are given in Figure 6b. By picking the appropriate values from the last two runs, the final model is given by

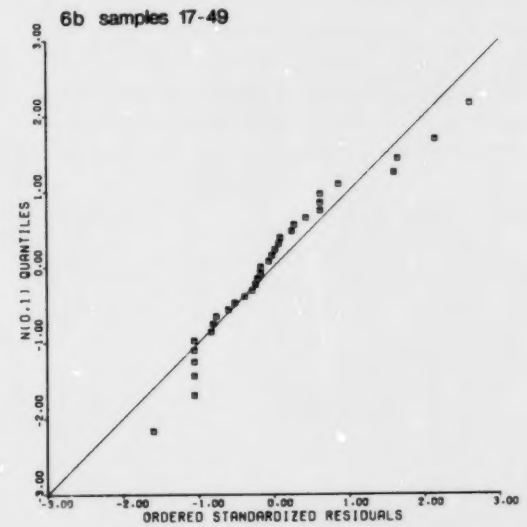
$$\hat{y}_i = 8.46 \qquad i = 1,2\ldots16$$

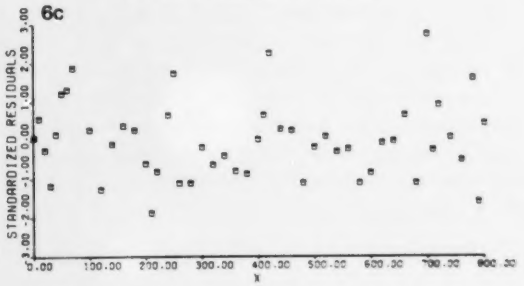
$$= 1.25 \qquad i = 17,18\ldots44$$

$$= 273 - 0.34\,x_i \qquad i = 45,46\ldots49$$









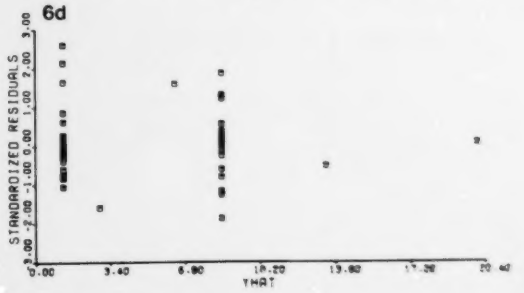


Figure 6. Plots constructed from runs 2 and 3 of Example 4.

with $\hat{\sigma}_1^2$=16.52, $\hat{\sigma}_2^2$=1.27 and $\hat{\sigma}_3^2$=2.62. Analogous to the calculation of $100R^2$ for two segments, the calculation for three segments leads to $100R^2$=75.5. Although some smaller peaks still appear in the residuals, the large changes have been located. Further, the observed significance level for the runs test based on the three-segment model is 0.20. The numbers required for calculating $100R^2$ and the significance level for the runs test were taken from one of the three program runs.

It is not always possible to obtain a satisfactory estimate of one of the points of change when more than one exists. This can occur if the changes are of similar magnitude. In this case, if segments of the data which contain only one change can be determined by examination of a scatter plot, a series of runs of the program will often provide satisfactory estimates.

DISCUSSION

The upper bounds for the degrees of the polynomials must be chosen carefully to prevent fitting small fluctuations in the data. As already emphasized, the fitted model should be examined for departures from the assumptions. The estimation of the point of change is sensitive to single points that are markedly different from surrounding points if the maximum value of the degree in one or both segments has been set too high. This will be obvious when the plots are examined.

The allowance for unequal variances in the two segments is an important feature in the procedure, since often when a system is disturbed the variability increases. Differences in the variability about the fitted curve in the two segments will not appear in the residual plots because standardized residuals are used. Changes in variability that do not take the simple form of constant variance within the segments can be detected from the residual plots and runs test.

The examples have not shown how the program is used for more than one pair of variables. If the three cases had been run separately in Example 4 without requesting plots, one run could then have been done for the three cases with plots requested. The same two columns of data would be specified for each case. A more usual situation would be to specify different pairs of columns. For Example 4, the columns might have corresponded to the three pairs: (1) depth and *Pinus-banksiana/resinosa* type, (2) depth and *Pinus strobus* type and (3) depth and the sum of *Pinus-banksiana/resinosa* type and *Pinus strobus* type.

ACKNOWLEDGMENTS

The algorithm was coded and the program tested by J. Hodson. Plot subroutines were originally written by D. Pateman and an earlier version of a similar program by M. Fellowes.

REFERENCES

Draper, N.R. and H. Smith. 1981. *Applied Regression Analysis*. 2nd ed., John Wiley and Sons, Inc., New York.

Esterby, S.R. and A.H. El-Shaarawi. 1980. Examples of inferences about the point of change in a regression model. Unpubl. rep. National Water Research Institute, Burlington, Ontario.

Esterby, S.R. and A.H. El-Shaarawi. 1981a. Inference about the point of change in a regression model. J. Royal Stat. Soc. Series C, 30(3): 277-285.

Esterby, S.R. and A.H. El-Shaarawi. 1981b. Likelihood inference about the point of change in a regression regime. J. Hydrol. 53: 17-30.

McAndrews, J.H. 1966. Postglacial history of prairie, savanna and forest in Northwestern Minnesota. Mem. Torrey Bot. Club, Vol. 22, No. 2, 72 pp.

Quandt, R.E. 1958. The estimation of parameters of a linear regression system obeying two separate regimes. J. Am. Stat. Assoc. 53: 873-880.

Robbins, J.A., D.N. Edgington and A.L.W. Kemp. 1978. Comparative ^{210}Pb, ^{137}Cs, and pollen geochronologies of sediments from Lakes Ontario and Erie. Quat. Res. 10(2): 256-278.

APPENDIX

Standardization of the independent variable by subtraction of the mean and division by the standard deviation, where the mean and standard deviation are calculated from all n values, has proven sufficient. The standardization which is analogous to the use of the correlation matrix in the calculations for a single regression function is to standardize each segment separately. However, this would increase the calculations substantially, since the samples in the segment change as the different possible values of n_1 are used. Thus the model that is fitted in one segment is

$$y_i = \beta_0 + \beta_1\left(\frac{x_i-\bar{x}}{s}\right) + \ldots + \beta_d\left(\frac{x_i-\bar{x}}{s}\right)^d$$

and to obtain $\theta_0, \theta_1 \ldots \theta_d$ which correspond to the model in terms of the original variables, $\underline{\theta} = (\theta_0, \theta_1 \ldots \theta_d)'$ is calculated from

$$\underline{\theta} = A\underline{\beta}^*$$

where $\underline{\beta}^* = \begin{bmatrix} \beta_0 \\ \beta_1/s \\ \ldots \\ \beta_d/s^d \end{bmatrix}$

and A is a (d+1) x (d+1) matrix which must be generated. The columns consist of terms of the binomial series $(1-\bar{x})^{j-1}$ for j = 1,2. . .d+1. The first j elements of the jth column of A are given by the terms of the series, in reverse order, and the last (d-j+1) elements are zero. For example if d=2, then

$$\begin{bmatrix} \theta_0 \\ \theta_1 \\ \theta_2 \end{bmatrix} = \begin{bmatrix} 1 & -x & x^2 \\ 0 & 1 & -2x \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1/s \\ \beta_2/s^2 \end{bmatrix}$$

The multiple correlation coefficient, R^2, is defined as the regression sum of squares divided by the total sum of squares (TSS). In the present model, the numerator of R^2 is not just the sum of the regression sums of squares calculated separately for each segment, since this does not involve all of the components of the model. To see this the total sum of squares can be decomposed as follows:

$$TSS = \sum_{i=1}^{n} (y_i - \bar{y})^2 = \sum_{i=1}^{n_1} (y_i - \bar{y})^2 + \sum_{i=n_1+1}^{n} (y_i - \bar{y})^2$$

$$= TSS_1 + TSS_2$$

But

$$TSS_1 = \sum_{i=1}^{n_1} (y_i - \bar{y}_1)^2 + \sum_{i=1}^{n_1} (\bar{y}_1 - \bar{y})^2$$

and the first term equals

$$\sum_{i=1}^{n_1} (y_i - \hat{y}_{1i})^2 + \sum_{i=1}^{n_1} (\hat{y}_{1i} - \bar{y}_1)^2$$

where the subscript 1 refers to a quantity calculated from the samples in segment one. These two terms correspond to the residual and regression sums of squares for segment one. By doing the same decomposition for segment two, recombining and grouping together the corresponding terms for both segments, it follows that

$$TSS = \{ResSS_1 + ResSS_2\} + \{RegSS_1 + RegSS^2\} + M$$

where $M = n_1(\bar{y}_1-\bar{y})^2 + (n-n_1)(\bar{y}_2-\bar{y})^2$ and ResSS and RegSS denote residual and regression sums of squares, respectively. Thus

$$100R^2 = 100\{TSS - (ResSS_1 + ResSS_2)\}/TSS$$

Environnement Canada Environment Canada

Programme d'estimation du point de transition et des degrés des polynômes d'approximation d'une fonction de régression

S.R. Esterby

ÉTUDE Nº 147, SÉRIE SCIENTIFIQUE

DIRECTION GÉNÉRALE DES EAUX INTÉRIEURES
INSTITUT NATIONAL DE RECHERCHES SUR LES EAUX
CENTRE CANADIEN DES EAUX INTÉRIEURES
BURLINGTON (ONTARIO) 1985

(Available in English on request)

© Ministre des Approvisionnements et Services Canada 1986

N° de cat. En36-502/147F

ISBN 0-662-93696-5

Table des matières

Page

RÉSUMÉ v

ABSTRACT v

INTRODUCTION 1

ALGORITHME 2

EXEMPLES 5
Exemple 1 5
Exemple 2 9
Exemple 3 11
Exemple 4 12

ANALYSE 13

REMERCIEMENTS 13

RÉFÉRENCES 13

ANNEXE 15

Tableaux

1. Choix de l'utilisateur pour chaque passage du programme 2
2. Différences selon l'hypothèse sur les variances 4
3. Choix pour les exemples présentés 5

Illustrations

Figure 1. Organigramme général du déroulement du programme 3
Figure 2. État produit par l'ordinateur lors du premier passage du programme pour l'analyse de l'échantillon KB, exemple 1 6
Figure 3. Graphiques produits par l'ordinateur pour l'analyse de l'échantillon KB, exemple 1 8
Figure 4. Représentation graphique des données avec les courbes de lissage et la fonction de vraisemblance relative, exemple 2 10
Figure 5. Représentation graphique des écarts normalisés, exemple 3 11
Figure 6. Graphiques tracés à partir des passages 2 et 3 du programme, exemple 4 12

Résumé

On décrit un programme informatique en Fortran qui utilise une méthode de calcul pour un modèle de régression polynomiale avec un point de transition. Les hypothèses posées pour le modèle sont les mêmes que pour un seul polynôme, mais on suppose que le polynôme, et peut-être la variance, changent en un point de la série de données. L'ordre des données simples doit aussi être connu. Le degré du polynôme dans chaque segment et le point de transition sont calculés par une méthode itérative qui comporte l'emploi de la fonction de vraisemblance marginale pour le paramètre du point de transition et un test faisant appel au changement des sommes des carrés des écarts résiduels avec une diminution du degré. Les hypothèses du modèle peuvent être examinées à l'aide des valeurs des écarts résiduels et de la fonction de vraisemblance. On donne des exemples qui démontrent l'utilité du programme et indiquent les résultats obtenus avec l'imprimante et le traceur.

Abstract

A Fortran computer program, which implements an estimation procedure for a polynomial regression model with a change point, is described. The assumptions of the model are analogous to those for a single polynomial but include the assumption that the polynomial, and possibly the variance, have changed at a point in the data set. The order of the data points must also be known. The degree of the polynomial in each segment and the point of change are estimated by an iterative procedure which involves the use of the marginal likelihood function for the change-point parameter and a test using the change in residual sums of squares with reduction of degree. The assumptions of the model can be examined using the residuals and the likelihood function. Examples are given to demonstrate the utility of the program and to show the output from both the printer and the plotter.

Programme d'estimation du point de transition et des degrés des polynômes d'approximation d'une fonction de régression

S.R. Esterby

INTRODUCTION

On emploie souvent les méthodes de l'analyse de régression pour obtenir des renseignements sur les relations qui lient diverses variables. Dans la plupart des applications, une relation de régression unique est ajustée à un certain ensemble de données; il arrive toutefois que cette relation change de forme au cours de la période de collecte des données. La complexité de l'analyse dépend de la connaissance qu'on a de la forme de cette relation ainsi que du point de transition. Lorsque le point de la transition d'une relation de régression à l'autre est connu, l'analyse consiste à approximer chacun des segments par la méthode des moindres carrés, en imposant éventuellement aux fonctions de régression la contrainte supplémentaire qu'elles doivent avoir la même valeur au point de transition. Lorsque ce dernier est inconnu, l'analyse se complique car il faudra alors estimer le point de transition et aussi les paramètres de régression. Nous nous intéresserons, dans le présent rapport, à cette dernière situation et décrirons un algorithme fondé sur une méthode proposée par Esterby et El-Shaarawi (1981a).

Cet algorithme a été établi pour le cas d'une seule variable dépendante. On suppose que chaque segment de l'ensemble de données considéré peut être approximé par une fonction de régression unique (en appliquant la méthode des moindres carrés) et qu'en outre la relation de régression s'est modifiée à un certain point qu'il reste à déterminer. Nous supposerons que la fonction de régression est un polynôme et que l'ensemble des observations est un ensemble ordonné (par exemple par un ordre naturel dans le temps ou dans l'espace). Soit (x_i, y_i) un ensemble de paires d'observations indépendantes, où l'indice i désigne l'ordre de l'observation. Le modèle consiste alors en une paire de polynômes de degrés respectifs p et q, la transition du polynôme de degré p au polynôme de degré q s'effectuant après l'observation pour laquelle on a $i=n_1$. On a donc

$$y_i = \sum_{j=0}^{p} \theta_{1j}\, x_i^j + e_{1i} \qquad i = 1,2\ldots n_1$$

et

$$y_i = \sum_{j=0}^{q} \theta_{2j}\, x_i^j + e_{2i} \qquad i = n_1+1, n_1+2\ldots n$$

où e_{1i} $(i=1,2\ldots n_1)$ et e_{2i} $(i=n_1+1, n_1+2\ldots n)$ sont deux suites de variables aléatoires indépendantes que l'on suppose respectivement $N(0, \sigma_1^2)$ et $N(0, \sigma_2^2)$. Les valeurs des paramètres $\underline{\theta}_1 = (\theta_{10}, \theta_{11}\ldots\theta_{1p})$, $\underline{\theta}_2 = (\theta_{20}, \theta_{21}\ldots\theta_{2q})$, σ_1^2, σ_2^2 et n_1 sont inconnues. Dans les deux sections qui suivent, ce modèle est désigné par la notation $M_u(p,q|n_1)$. Puisque p et q sont donnés, on cherche à estimer la valeur du point de transition n_1 à partir de la fonction de vraisemblance marginale de n_1. Dans le cas où p et q sont également inconnus, on obtient une estimation de ces deux paramètres par itération. La méthode consiste à exploiter conjointement la fonction de vraisemblance marginale de n_1 et un test analogue au test de réduction du degré d'un polynôme unique. On suppose que le degré de chaque polynôme est borné supérieurement. Soit (p^*, q^*) la paire des bornes supérieures des degrés des deux polynômes; à chaque étape du processus d'itération, le degré total $(p + q)$ est réduit d'une unité, jusqu'à ce qu'aucune réduction ne soit plus possible ou que $(p,q) = (0,0)$. Le cas où $\sigma_1^2 = \sigma_2^2$ est également envisagé; le modèle correspondant est désigné par la notation $M_e(p,q|n_1)$, où les indices e et u servent à distinguer les modèles correspondant respectivement à l'inégalité des variances. Il existe en effet des différences marquées entre les fonctions de vraisemblance marginale et les méthodes d'estimation de p et q selon que les variances sont égales ou non. On trouvera une description détaillée de ces méthodes dans Esterby et El-Shaarawi (1981a) ainsi qu'une description plus détaillée des fonctions de vraisemblance dans Esterby et El-Shaarawi (1981b).

Les méthodes de l'analyse des écarts résiduels, qui sont applicables au cas où l'ensemble de données considéré peut être approximé par une fonction unique, peuvent être transposées ici : il suffit en effet d'analyser l'ensemble des

écarts résiduels correspondant à chaque segment. En outre, l'étude de la fonction de vraisemblance permet de confirmer ou d'infirmer l'hypothèse de l'existence d'un point de transition unique.

Le but principal de la méthode décrite ici consiste à estimer le point où la relation de régression change de forme dans l'hypothèse de l'existence d'un point de transition. Aucune contrainte n'est imposée aux valeurs des fonctions de régression en ce point. Selon l'objectif visé, il est possible que des méthodes différentes doivent être employées. Si l'on désire par exemple que la courbe qui approxime l'ensemble de données considéré soit une courbe lisse, on peut dans de nombreux cas approximer les données au moyen d'une spline cubique ou d'un polynôme de degré supérieur.

On trouvera une description de l'algorithme à la section suivante ainsi que des exemples d'utilisation du programme dans la troisième section. La séquence en Fortran n'a pas été listée parce que le programme est très long. On peut se procurer une copie de ce dernier ainsi que des jeux d'essai en écrivant à l'auteur. Le programme, rédigé en Fortran V, comporte un certain nombre de sous-programmes IMSL et est exécuté sur un CDC Cyber 171 tournant sous NOS. Les programmes de traçage ont été rédigés pour un traceur CALCOMP 1036. Nous décrirons ici les principales caractéristiques du programme sans pour cela reprendre tout le développement de la méthode. Nous illustrerons son utilisation par des exemples, de façon que le lecteur qui a besoin d'un programme de cette nature puisse se faire une idée de l'utilité de celui que nous avons mis au point pour l'application particulière qu'il a en vue.

ALGORITHME

De nombreuses fonctions ont été intégrées au programme pour permettre à l'utilisateur d'analyser simplement et efficacement une ou plusieurs paires de variables en un seul passage, pourvu que toutes ces variables soient stockées dans le même fichier. On trouvera dans le tableau 1 la liste des options que l'utilisateur doit préciser. Le programme est en mesure d'estimer la valeur du point de transition n_1 pourvu qu'on lui fournisse les degrés des polynômes ou d'estimer les degrés des polynômes ainsi que la valeur du point de transition pourvu qu'on lui indique la borne supérieure du degré du polynôme dans chaque segment. Pour chaque paire de variables, l'utilisateur peut préciser soit la valeur du couple de degrés (p,q) soit cinq valeurs pour les couples de bornes supérieures (p^*,q^*). Pour la paire (p,q) ou pour chaque paire (p^*,q^*), l'utilisateur doit préciser si les variances sont supposées égales ou non et si des graphiques doivent être tracés ou non; il doit également préciser les deux niveaux de signification critique correspondant à chaque test d'hypothèse. Les graphiques tracés varient selon que le domaine de la variable indépendante est ordonné de façon monotone (croissante ou décroissante) ou non.

On trouvera représentées à la figure 1 les principales caractéristiques du programme; cette figure montre également l'effet des options choisies par l'utilisateur sur les parties du programme qui sont utilisées. Pour comprendre les détails des calculs qui sont effectués à chaque étape de traitement de l'organigramme, il faut lire l'article de Esterby et El-Shaarawi (1981a). Nous allons toutefois faire ici un certain nombre d'observations qui aideront à clarifier la

Tableau 1. Choix de l'utilisateur pour chaque passage du programme

Niveau d'application	Choix
	Le nombre de paires de variables
Pour chaque paire de variables	La forme selon laquelle les valeurs des variables x et y sont lues
	Le titre
	Si oui ou non l'ensemble des valeurs de la variable indépendante est ordonné
	Soit 1) la valeur de (p, q) si elle est connue ou
	2) le nombre de paires de bornes supérieures n_{pq} des degrés des polynômes et leurs valeurs (p^*, q^*)
	Si tous les échantillons doivent être utilisés ou si certains (et lesquels) doivent être omis
	Les unités adoptées pour les axes de représentation graphique
Pour la paire (p,q) ou pour chaque paire (p^*,q^*)	Si les variances sont égales, inégales ou s'il faut effectuer un test pour en déterminer l'égalité ou l'inégalité
	Si des graphiques doivent être produits
	Le niveau de signification critique pour le test d'égalité des variances
	Le niveau de signification critique pour le test de réduction du degré

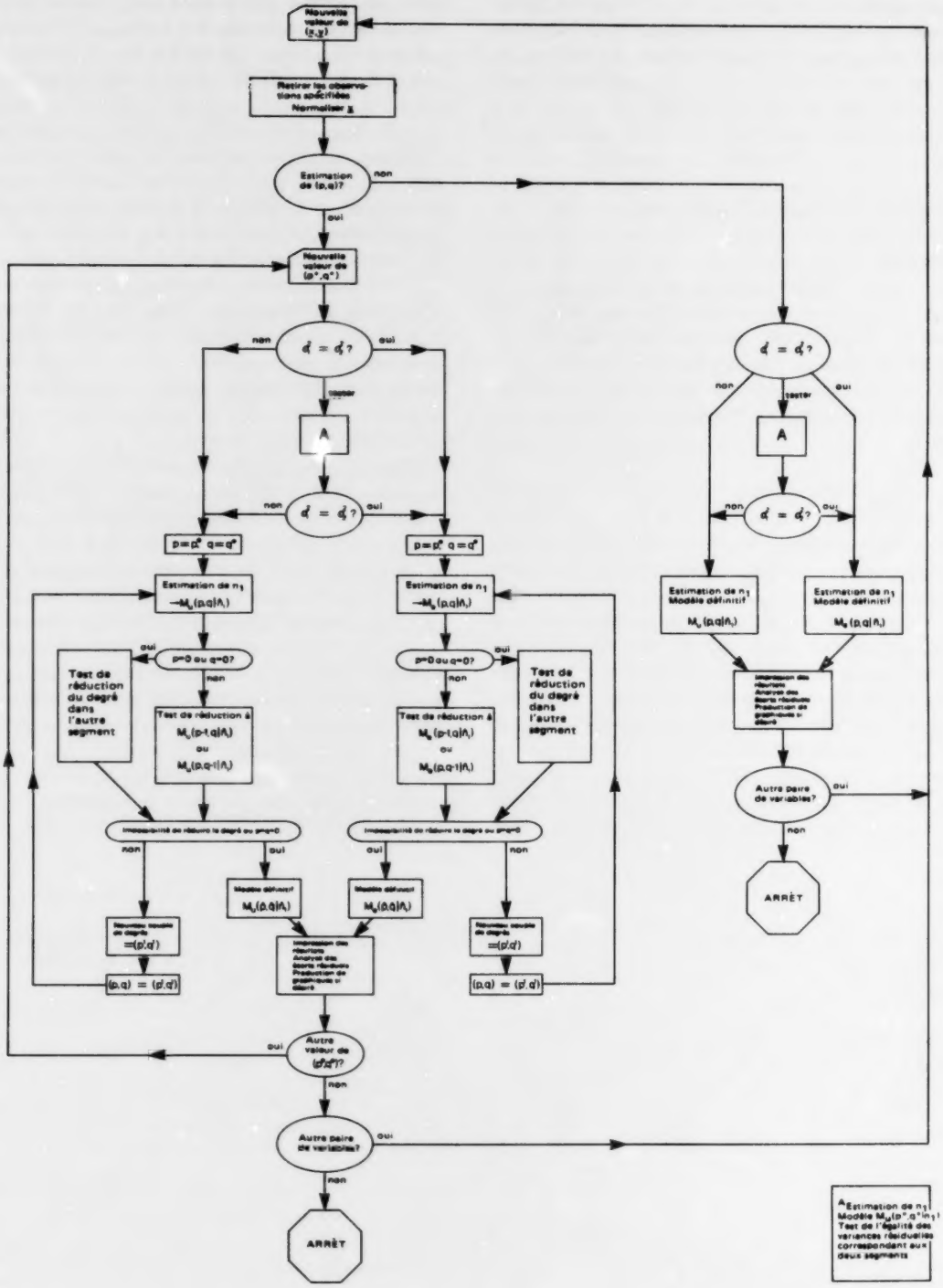


Figure 1. Organigramme général du déroulement du programme.

structure logique du programme. La méthode d'itération qui permet d'estimer p et q commence par fixer à ces deux paramètres des valeurs particulières puis, à partir de ces valeurs, attribue une valeur $\hat{n}_1$ à n_1. En supposant ensuite que n_1 prend cette valeur $\hat{n}_1$, on effectue un test pour déterminer si le degré total (p+q) ne pourrait pas être réduit d'une unité. Dans l'affirmative, on obtient à partir de (p+q-1) une nouvelle estimation de n_1. Les deux suites de la figure 1, qui correspondent respectivement au cas où les variances sont égales et inégales, sont expliquées avec plus de détails dans le tableau 2, où les différences attribuables à l'égalité ou à l'inégalité des variances sont précisées. La valeur estimée $\hat{n}_1$ de n_1 est la valeur de ce paramètre qui maximise la fonction de vraisemblance marginale; on détermine cette valeur en calculant les valeurs de la fonction de vraisemblance marginale pour les valeurs de n_1 qui correspondent aux valeurs de l'indice qui apparaissent dans le tableau 2.

Plusieurs des étapes de la figure 1 n'apparaissent pas dans la méthode décrite dans l'article de 1981a par Esterby et El-Shaarawi. Pour déterminer si les variances sont égales ou non (voir A à la figure 1), on commence par faire l'hypothèse que les variances en question sont inégales et on obtient ensuite une estimation du point de transition à partir de la fonction de vraisemblance marginale. On applique ensuite un test F standard pour vérifier l'hypothèse de l'égalité des variances résiduelles correspondant aux segments 1 et 2. Lors de l'estimation des degrés p et q, lorsque l'un de ces degrés s'annule dans le cadre du processus d'itération, il suffit d'appliquer la méthode de réduction à l'autre segment. On se trouve alors dans le cas d'un test de réduction de degré correspondant à l'ajustement d'un polynôme unique à un ensemble de données. Dans ce cas, on peut utiliser directement la valeur α du niveau de signification critique, alors que dans le cas où le degré peut être réduit d'une unité de deux façons, on doit employer $\sqrt{\alpha}$.

La mise en oeuvre de l'algorithme que nous avons mis au point comporte également plusieurs étapes additionnelles. Pour permettre l'inversion de certaines matrices, on commence par normaliser la variable indépendante. Il faut ensuite effectuer des calculs additionnels pour exprimer les paramètres de régression en fonction de la variable indépendante d'origine, connaissant la relation qui lie ces paramètres et la variable normalisée. On trouvera dans l'annexe les formules utilisées à cette fin. Dans le cas où les variances sont égales et où ni p ni q n'est nul, le dénominateur de la fonction test de réduction du degré comporte la somme des sommes des carrés des écarts résiduels correspondant aux deux segments. Les bornes inférieure et supérieure de n_1 peuvent donc être choisies respectivement égales à p+1 et n-(q+1), de sorte que le dégré de liberté de la somme des carrés des écarts résiduels correspondant à un segment est nul. Toutefois, lorsque cela se produit en même temps que le degré du polynôme qui correspond à l'autre segment est nul, on obtient pour la fonction test un dénominateur à 0 degré de liberté, ce qui est impossible. La solution que nous avons adoptée pour résoudre ce problème consiste à modifier la limite correspondante pour n_1 de façon que le nombre d'observations soit toujours supérieur d'une unité au nombre de paramètres à estimer et à conserver cette relation valide pour la totalité du processus d'itération. Dans le cas où les variances sont égales et où le nombre de paramètres est égal au nombre d'observations dans un segment, on ne calcule pas la somme des carrés des écarts

Tableau 2. Différences selon l'hypothèse sur les variances

Caractéristiques	Variances égales	Variances inégales
Valeurs admissibles de n_1	p+1. . .n-(q+1)	p+2. . .n-(q+2)
Fonction de vraisemblance marginale de n_1, $L(n_1)$	$\{\nu_1\hat{\sigma}_1{}^2+\nu_2\hat{\sigma}_2{}^2\}^{-1/2(\nu_1+\nu_2-1)}$	$\frac{\Gamma(1/2\nu_1)\ \Gamma(1/2\nu_2)}{\nu_1{}^{1/2\nu_1}\ \nu_2{}^{1/2\nu_2}}\ \hat{\sigma}_1{}^{-(\nu_1-1)}\ \hat{\sigma}_2{}^{-(\nu_2-1)}$
Test de réduction du degré, p et q non nuls	On regroupe les sommes des carrés des écarts résiduels correspondant aux deux segments pour estimer la variance globale.	On calcule séparément les sommes des carrés des écarts résiduels correspondant à chaque segment pour estimer les variances.
Écarts résiduels normalisés	On divise les écarts résiduels correspondant aux deux segments par $\hat{\sigma}$ dont la valeur a été déterminée à partir du regroupement des sommes des carrés des écarts résiduels	On normalise tout d'abord les écarts résiduels en les divisant par $\hat{\sigma}_1$ ou $\hat{\sigma}_2$ puis on les regroupe.

Nota: $\nu_1 = n_1-(p+1)$, $\nu_2 = (n-n_1) - (q+1)$, $\hat{\sigma}_1{}^2 = \sum_{i=1}^{n_1} (y_i - \sum_{j=0}^{p} \hat{\theta}_{1j}\ x_i^j)^2/\nu_1$,

$\hat{\sigma}_2{}^2 = \sum_{i=n_1+1}^{n} (y_i - \sum_{j=0}^{q} \hat{\theta}_{2j}\ x_i^j)^2/\nu_2$ et $\Gamma(z)$ désigne la fonction gamma de z.

résiduels; on lui attribue plutôt la valeur théorique 0. La somme des carrés des écarts résiduels prend également cette valeur lorsque la variable dépendante est constante dans un segment. Dans le cas où les variances ne sont pas égales, l'attribution de la valeur zéro à la somme des carrés des écarts résiduels ne constitue pas une solution satisfaisante, puisque le logarithme de la somme des carrés des écarts résiduels correspondant à chaque segment intervient dans la fonction de vraisemblance marginale. Dans le cas où la variable dépendante prend une valeur constante dans un segment, on attribue arbitrairement à la somme des carrés des écarts résiduels correspondant à ce segment la valeur 0.000 000 1. Le cas où la variable dépendante a une valeur constante est donc traité de la même façon, que les variances soient égales ou non. L'utilisateur peut contourner ce choix arbitraire en indiquant tout simplement au programme d'omettre la valeur finale de la constante.

Les résultats ci-après, notamment, sont imprimés : un état récapitulatif des étapes de l'estimation de p et q, et, dans le cas du modèle définitif, la fonction de vraisemblance marginale relative de n_1, les paramètres de régression, $\hat{\underline{\theta}}_1$ et $\hat{\underline{\theta}}_2$, la valeur estimée de $\underline{y}$, $\hat{\underline{y}}$, ainsi que les écarts résiduels $\hat{\underline{e}}=\underline{y}-\hat{\underline{y}}$. La somme des carrés des écarts résiduels, le carré moyen des écarts, la somme quadratique totale corrigée ainsi que R^2 (cette valeur est définie en annexe) sont également imprimés. L'analyse des écarts résiduels (Draper et Smith, 1981) s'effectue essentiellement par une méthode graphique : on trace un graphique Q-Q pour vérifier l'hypothèse de normalité ainsi que des graphiques de représentation des écarts résiduels normalisés en fonction de $\underline{x}$ et $\hat{\underline{y}}$. On applique également un test de suites aux écarts résiduels.

EXEMPLES

Des exemples visant à évaluer l'efficacité de la méthode décrite ici ont déjà été présentés dans des publications antérieures (Esterby et El-Shaarawi, 1980, 1981a,b). Les exemples que nous présentons ici visent surtout à illustrer le rôle du programme sur ordinateur ainsi que ses possibilités. L'exemple 1 est traité en détail et l'état ainsi que les graphiques produits par l'ordinateur sont présentés ci-dessous. Dans les trois autres exemples, nous n'avons présenté que les graphiques qui illustraient certaines caractéristiques particulières.

Exemple 1

Dans certaines conditions, la distribution verticale du pollen de type *Ambrosia* dans une carotte sédimentaire peut servir de moyen de datation. Les hypothèses de l'existence d'un point de transition unique et de la représentabilité de la relation de régression qui lie la concentration et la profondeur par une fonction polynômiale sont plausibles. L'ensemble d'échantillons utilisé ici est décrit par Robbins, Edgington et Kemp (1978); c'est un ensemble

Tableau 3. Choix pour les exemples présentés

	Exemple						
Option	1a	1b	2	3	4a	4b	4c
Domaine de x ordonné	oui	oui	non	oui	oui	oui	oui
(p,q) ou		(1,0)	(1,1)	(0,0)			(0,1)
(p*,q*)	(1,1)				(3,3)	(1,1)	
Échantillons utilisés	tous	tous	tous	tous	tous	1-44	17-49
Variances	tester	tester	tester	égales	tester	tester	tester
α, variances	0.05	0.05	0.05	0.05	0.05	0.05	0.05
α, degré	0.01	0.01	0.01	0.01	0.01	0.01	0.01
Graphiques	oui	oui	oui	oui	non	oui	oui
x borne inférieure	0.0	0.0	0.0	0.0		0.0	0.0
borne supérieure	68.4	68.4	18.0	100.0		800.0	800.0
accroissement	7.6	7.6	2.25	10.0		100.0	100.0
y borne inférieure	0.0	0.0	0.0	-3.3		0.0	0.0
borne supérieure	47.5	47.5	20.04	3.3		20.4	20.4
accroissement	9.5	9.5	3.34	1.1		3.4	3.4

Nota : Les lettres a,b,c désignent différents passages du programme correspondant à un même exemple; ainsi, 1a désigne le premier passage du programme pour l'exemple 1.

```
                                   S.E.G.R.E.G

DATA SET  1

FORMAT FOR READING X AND Y: (TR6,F6.2,TL11,F5.1)
CORE KB AMBROSIA
X IS IN EITHER INCREASING OR DECREASING ORDER
P AND Q WILL BE ESTIMATED
 23 PAIRS OF (X,Y) VALUES READ:
             .500     46.600
            1.500     31.900
            3.000     37.500
            5.000     30.600
            7.000     34.400
            9.000     12.900
           11.000     30.300
           13.000     24.300
           15.000     20.400
           17.000     17.400
           19.000     27.800
           21.000     24.800
           23.000     12.900
           25.000     14.430
           27.000     11.400
           29.000      6.000
           35.000      6.400
           42.500      2.100
           47.500      2.600
           52.500      1.330
           57.500      2.000
           62.500      4.900
           67.500      1.200

DATA IS NOT MANIPULATED
MEAN OF X VALUES =   25.69565
STANDARD DEVIATION OF X VALUES =   20.36359
  0 ADJACENT EQUAL VALUES AT BEGINNING OF Y VECTOR
  0 ADJACENT EQUAL VALUES AT END OF Y VECTOR
PLOT SPECIFICATIONS
XORIGIN=     0.000    XLIMIT=    68.400    XSTEP=     7.600    YORIGIN=     0.000    YLIMIT=    47.500    YSTEP=     9.500

PAIR 1
------
MAXIMUM P = 1     MAXIMUM Q = 1
EQUALITY OF VARIANCES TO BE TESTED
PLOTS WILL BE PRODUCED
SIGNIFICANCE LEVEL FOR EQUALITY OF VARIANCES TEST =  .05000
SIGNIFICANCE LEVEL FOR DETERMINATION OF DEGREES OF POLYNOMIALS =  .01000

TEST FOR EQUALITY OF VARIANCES
------------------------------
           DEGREES OF        RESIDUAL          RESIDUAL
SEGMENT      FREEDOM            SS                MS
---------------------------------------------------------
   1           13       5.3674925157      .4128840397
   2            6        .1920299883      .0320048480
F RATIO =  12.90067
SIGNIFICANCE LEVEL =       .00246486
VARIANCES ARE UNEQUAL

ESTIMATION OF P AND Q, UNEQUAL VARIANCES
----------------------------------------
                    MODEL 1                 MODEL 2        INCREASE    F                  OBS        MAX        THEOR
       N1MAT       P        Q             P         Q       IN RSS   RATIO        DF       S.L.       S.L.       S.L.     CONCLUDE
----------------------------------------------------------------------------------------------------------------------------------
        15         1        1             0         1     9.496601 23.00065      1,13     .00035
                                          1         0      .122858  3.83874      1, 6     .09779     .09779     .10000  ADEQUATE
THE THEORETICAL SIGNIGICANCE LEVEL IS THE SQUARE ROOT OF SPECIFIED SIGNIGICANCE LEVEL IF THERE ARE TWO POSSIBLE WAYS
TO REDUCE THE DEGREE; OTHERWISE IT IS THE SPECIFIED SIGNIFICANCE LEVEL.

TABLE OF RESIDUAL SUMS OF SQUARES AND MEAN SQUARES CALCULATED DURING THE ESTIMATION OF P AND Q
----------------------------------------------------------------------------------------------
  N1MAT     P      Q           RSS1            RSS2            RMS1            RMS2
---------------------------------------------------------------------------------------
   15       1      1    5.3674925157    .1920299883     .4128840397     .0320048480
   15       0      0   14.8640933333    .3148875958    1.0617209524     .3449839286
```

Figure 2. État produit par l'ordinateur lors du premier passage du programme pour l'analyse de l'échantillon KB, exemple 1.

P = 1   Q = 1

N1	RESIDUAL SUMS OF SQUARES SEGMENT 1	RESIDUAL SUMS OF SQUARES SEGMENT 2	RESIDUAL MEAN SQUARES SEGMENT 1	RESIDUAL MEAN SQUARES SEGMENT 2	LIKELIHOOD FUNCTION (RELATIVE)
3	.0047605266	6.4240926088	.0047605266	.3568940338	.0135801626
4	.8095717292	6.2467128966	.4047858646	.3674536998	.0066384656
5	1.1529130989	5.1183713472	.3843043663	.3198982092	.0159189212
6	2.1584752322	4.1290101579	.5396188081	.2752673439	.0220677984
7	3.4917569930	3.4508413604	.6983513986	.2464886686	.0177281441
8	3.6681152344	3.2788882589	.6113525391	.2522221738	.0126684386
9	3.6960207362	3.2537018916	.5280029623	.2711418243	.0077257073
10	3.7031005032	3.2512143972	.4628875629	.2955649452	.0055890927
11	4.8302442976	2.0412903486	.5366938108	.2041290349	.0129160538
12	5.2489541575	.7732390644	.5248954158	.0859154516	.2678507197
13	5.3612703328	.7256368872	.4873882121	.0907046109	.1376431183
14	5.3634444412	.4193465694	.4469537034	.0599066528	.3402528632
15	5.3674925157	.1920293883	.4128840397	.0320048980	1.0000000000
16	5.4944584380	.1835927918	.3924613170	.0367185584	.3119150932
17	5.5683852142	.0903771429	.3712256809	.0225942857	.3174742161
18	5.7461140819	.0903600000	.3591321301	.0301200000	.0052934910
19	6.2395189234	.0871200000	.3670305249	.0435600000	.0210392022
20	6.7739365206	.0726000000	.3763298067	.0726000000	.0686527419

N1HAT	RESS1	RESS2	RMS1	RMS2
15	5.3674925157	.1920293883	.4128840397	.0320048980

I	Y=F(THETA)	Y=F(BETA)
1	36.94	36.94
N	1.50	1.50

ORDER OF OBSERVATION	OBSERVED X	OBSERVED Y	ESTIMATED Y	RESIDUAL	ORDERED STANDARDIZED RESIDUAL	STANDARD NORMAL QUANTILE
1	.5000000000	46.6000000000	36.9382360283	9.6617639717	-2.4937503539	-2.0190862006
2	1.5000000000	31.9000000000	35.9953677411	-4.0953677411	-1.0519682562	-1.5123895861
3	3.0000000000	37.5000000000	34.5810553103	2.9189446897	-.9447605475	-1.2334945620
4	5.0000000000	30.0000000000	32.6953287359	-2.6953287359	-.6373512724	-1.0271542729
5	7.0000000000	34.4000000000	30.8095921615	3.5904078385	-.6195304061	-.8572543202
6	9.0000000000	12.9000000000	28.9238555871	-16.0238555871	-.4952872339	-.7091025836
7	11.0000000000	30.3000000000	27.0381190127	3.2618809873	-.4461285360	-.5751092013
8	13.0000000000	24.3000000000	25.1523824383	-.8523824383	-.4394449119	-.4507442126
9	15.0000000000	20.4000000000	23.2666458639	-2.8666458639	-.4194668975	-.3330050902
10	17.0000000000	17.4000000000	21.3809092895	-3.9809092895	-.2762946346	-.2197229158
11	19.0000000000	27.0000000000	19.4951727151	7.5048272849	-.1690769659	-.1092004817
12	21.0000000000	24.0000000000	17.6094361407	6.3905638593	-.1326540292	0.0000000000
13	23.0000000000	12.9000000000	15.7236995663	-2.8236995663	-.0859415399	.1092004817
14	25.0000000000	14.4000000000	13.8379629919	.5620370081	.1874683362	.2197229158
15	27.0000000000	11.4000000000	11.9522264174	-.5522264174	.3796095133	.3330050902
16	29.0000000000	6.0000000000	5.3208817054	.6791182946	.4542661019	.4507442126
17	35.0000000000	6.4000000000	4.7258056445	1.6741943555	.5076379291	.5751092013
18	42.5000000000	2.1000000000	3.9819605685	-1.8819605685	.5547656959	.7091025836
19	47.5000000000	2.6000000000	3.4860638511	-.8860638511	.9358322093	.8572543202
20	52.5000000000	1.3000000000	2.9901671337	-1.6901671337	.9945465871	1.0271542729
21	57.5000000000	2.0000000000	2.4942704163	-.4942704163	1.1679564632	1.2334945620
22	62.5000000000	4.9000000000	1.9983736990	2.9016263010	1.5036346270	1.5123895861
23	67.5000000000	1.2000000000	1.5024769816	-.3024769816	1.6219355555	2.0190862006

SUM OF RESIDUALS, SEGMENT 1   .0000000000

SUM OF RESIDUALS, SEGMENT 2   .0000000000

TEST FOR RUNS IN RESIDUALS

NUMBER OF RUNS	16
NUMBER OF POSITIVE SIGNS	10
NUMBER OF NEGATIVE SIGNS	13
MEAN	12.3
STANDARD DEVIATION	2.30
N(0,1) VARIATE	1.82
PROBABILITY LEVEL	.97

FINAL MODEL

ALL RESIDUAL SUMS OF SQUARES AND RESIDUAL MEAN SQUARES WERE DIVIDED BY 100.

POINT OF CHANGE PARAMETER = 15

DEGREE OF POLYNOMIAL, SEGMENT 1 = 1

DEGREE OF POLYNOMIAL, SEGMENT 2 = 1

REGRESSION PARAMETERS

A. TRANSFORMED X						
SEGMENT 1	13.18285	-19.20019	0.00000	0.00000	0.00000	0.00000
SEGMENT 2	5.64866	-2.01965	0.00000	0.00000	0.00000	0.00000
B. UNTRANSFORMED X						
SEGMENT 1	37.40967	-.94287	0.00000	0.00000	0.00000	0.00000
SEGMENT 2	8.19708	-.09918	0.00000	0.00000	0.00000	0.00000

100 R SQUARED = 86.02406

TOTAL SUM OF SQUARES = 3977.92435

Figure 2. Suite.

de 23 couples dont le premier élément est une concentration de pollen de type *Ambrosia* (grains·g^{-1} de sédiment sec) et le deuxième élément la profondeur (cm) à laquelle on constate cette concentration; la carotte sédimentaire en question provient du lac Ontario. On s'est servi du milieu de l'intervalle d'échantillon pour indiquer la profondeur de cet échantillon. L'ensemble des échantillons est ordonné verticalement de haut en bas, l'échantillon prélevé en surface se voyant attribuer le rang 1. Les diverses options choisies sont indiquées dans le tableau 3 et l'état ainsi que les graphiques produits par l'ordinateur apparaissent dans les figures 2 et 3.

L'état produit par l'ordinateur comporte suffisamment de rubriques pour être facile à comprendre (figure 2). La conclusion qui découle du test portant sur l'égalité des variances s'applique à tous les résultats successifs. La fonction de vraisemblance relative, la somme des carrés des écarts résiduels, le carré moyen des écarts, la courbe de lissage ainsi que les écarts résiduels sont imprimés uniquement pour le modèle définitif. La dernière partie de l'état présente un récapitulatif décrivant le modèle définitif. Le facteur qui y apparaît doit être appliqué à toutes les sommes de carrés d'écarts résiduels ainsi qu'à tous les carrés moyens qui figurent dans les résultats précédents.

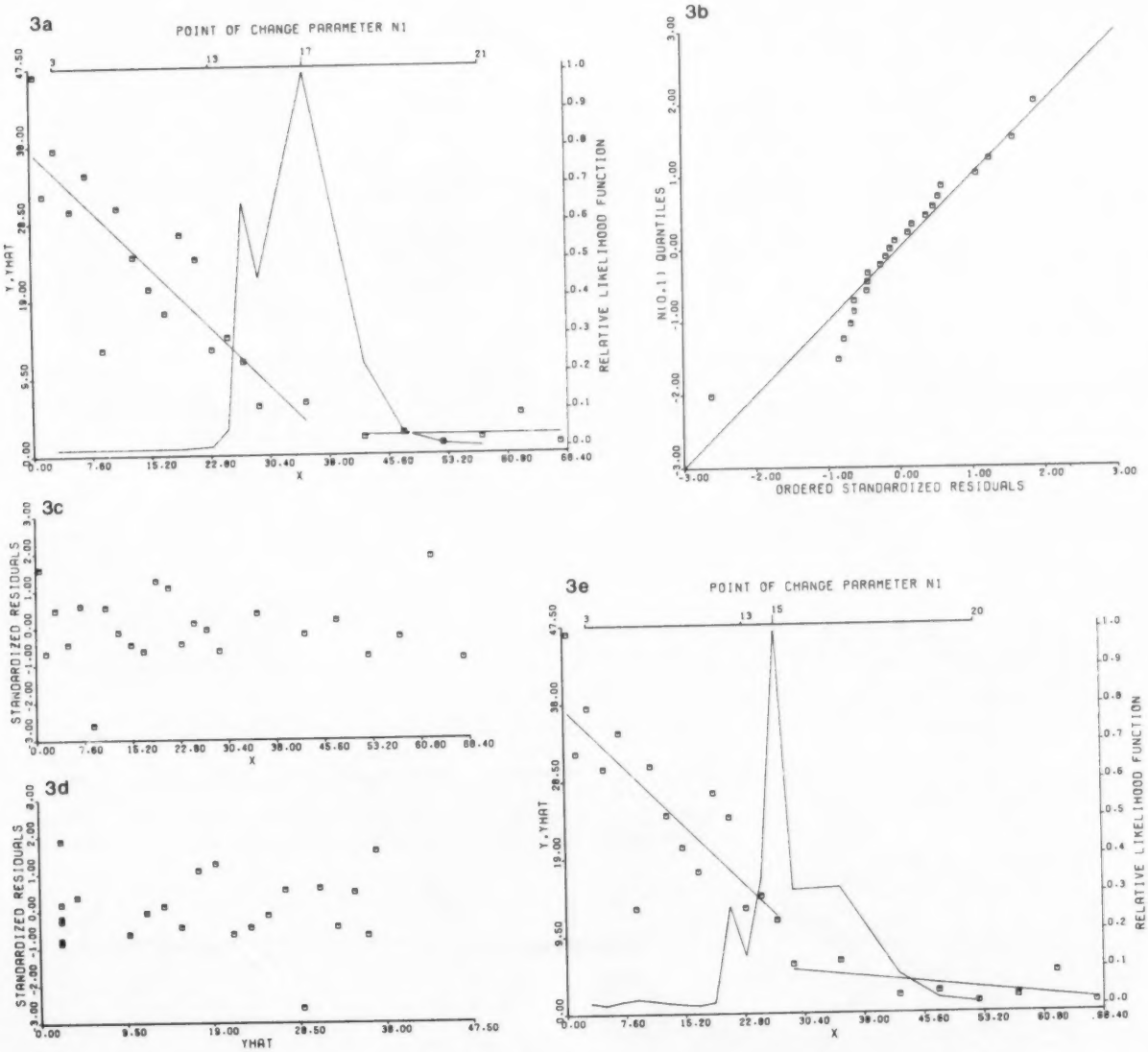


Figure 3. Graphiques produits par l'ordinateur pour l'analyse de l'échantillon KB, exemple 1.

Le modèle définitif est le modèle $M_u(1,1|15)$; il s'agit donc de deux droites de pentes différentes, une pour les échantillons 1 à 15 et une autre pour les échantillons 16 à 23; les variances qui correspondent à ces deux segments de droite sont différentes. Les variations de la concentration du pollen étant en général décrites de bas en haut, le modèle ci-dessus correspond à une augmentation du degré de concentration au taux de 0.10 grain$\cdot g^{-1} \cdot cm^{-1}$ du dernier échantillon du bas jusqu'à l'échantillon 16, puis au taux de 0.94 grain$\cdot g^{-1} \cdot cm^{-1}$ de l'échantillon 15 à l'échantillon prélevé à la surface. Des écarts plus grands par rapport à la moyenne ont également été constatés pour les 15 échantillons de la partie supérieure; on a en effet $\hat{\sigma}_1^2$=41.29 et $\hat{\sigma}_2^2$=3.20. Dans la figure 2, les taux et les variances sont exprimés en fonction de la valeur non transformée de x et de la somme des carrés des écarts résiduels correspondant à chaque segment. Le point de transition, l'échantillon 15, est le point où la fonction de vraisemblance relative atteint sa valeur maximum de 1. On peut se convaincre que l'échantillon 15 constitue bien le point de transition en étudiant la fonction de vraisemblance relative $R(n_1)$. Dans ce cas particulier, la valeur de la fonction de vraisemblance relative est $\geq$0.10 pour tous les échantillons qui appartiennent à l'intervalle fermé [12,17]. En examinant, par exemple, l'intervalle de vraisemblance relative qui correspond à la valeur 0.10, on obtient une certaine mesure du degré d'incertitude associé à l'estimation considérée.

Dans le cas où les valeurs de la variable indépendante sont ordonnées de façon croissante ou décroissante, quatre graphiques sont produits par l'ordinateur. La figure 3a représente les données, les courbes de lissage ainsi que la fonction de vraisemblance relative. Les données et les courbes de lissage sont repérées par rapport à l'axe vertical gauche et l'axe horizontal inférieur alors que la fonction de vraisemblance relative pour le paramètre du point de transition, n_1, est repérée par rapport à l'axe vertical droit et l'axe horizontal supérieur. On observera que seules les valeurs entières de n_1 donnent lieu à une valeur non nulle de la fonction de vraisemblance relative; toutefois, pour faciliter la lecture du graphique, nous avons joint les valeurs de cette dernière. Les graphiques de représentation des écarts résiduels dans les figures 3b, c et d ne font apparaître aucune divergence significative par rapport aux hypothèses retenues par le modèle. L'écart résiduel normalisé dont la valeur absolue est la plus grande est celui qui correspond à l'échantillon 6 (valeur -2.5) et, sous des hypothèses de normalité, on s'attend à ce qu'environ 98 % des écarts résiduels appartiennent à l'intervalle [-2.5, 2.5]. En outre, le niveau de signification observé (0.97) lors du test des suites (figure 2) ne vient pas contredire l'hypothèse de l'indépendance admise par le modèle.

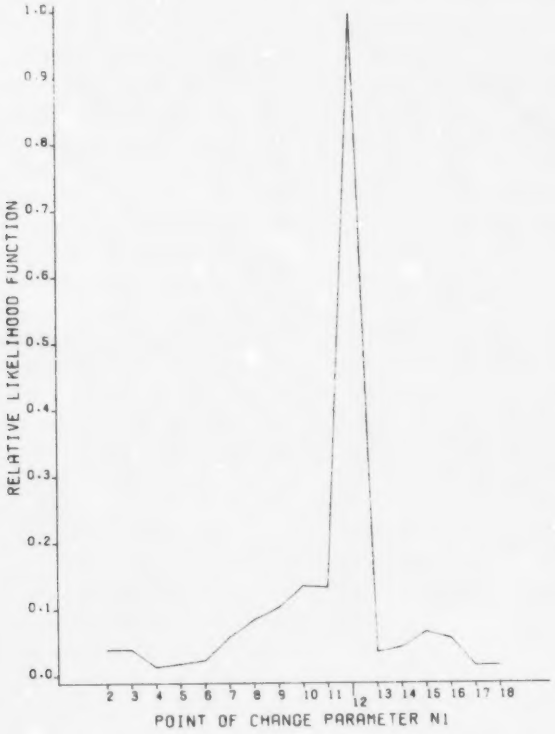
La connaissance générale de l'intervalle de vraisemblance auquel appartient le point de transition nous serait fort utile pour dater notre échantillon par la méthode de la distribution verticale du pollen de type *Ambrosia*. Or, selon Robbins et coll., on ne possède pas de définition précise du point de transition appliquée au pollen de type *Ambrosia*. Il est intéressant de noter que si cela désigne le point à partir duquel la concentration du pollen se met à augmenter à un taux accru, cela implique que q=0 pour le modèle décrit ici. Lorsqu'on attribue au couple (p,q) la valeur (1,0), on obtient le modèle M_u (1,0|17) (figure 3e) pour lequel l'intervalle de vraisemblance relative à 0.10 est l'intervalle [15, 18]. L'analyse des écarts résiduels n'a donné lieu à aucune différence marquée et on a $100R^2$=86.0 et 86.8 pour les modèles $M_u(1,1|15)$ et $M_u(1,0|17)$ respectivement. Dans Robbins et coll., le point de transition, dans le cas du pollen de type *Ambrosia*, a été fixé à l'échantillon 15 après analyse des données. Ce résultat correspond bien au premier modèle (deux droites) auquel nous sommes arrivés grâce à la méthode décrite ici. La comparaison des résultats obtenus par notre méthode et par la méthode d'inspection visuelle respectivement devrait aider l'analyste à mettre au point une méthode plus efficace pour la détermination du point où la fonction de répartition du pollen de type *Ambrosia* change de forme.

En général, la décision d'attribuer une valeur particulière au couple (p^*,q^*) ou de supposer le couple (p,q) connu dépend de la situation particulière dans laquelle on se trouve. Dans l'exemple du pollen de type *Ambrosia*, en attribuant une valeur supérieure à (p^*,q^*), on aurait obtenu un modèle mieux ajusté aux faibles variations des données mais qui n'aurait pas tenu compte de l'existence d'un point de transition à partir duquel le degré de concentration de ce type de pollen augmente dans les dépôts de sédiments du lac. Toutefois, à partir de l'expérience que l'auteur a acquise dans l'analyse d'autres échantillons, il est clair qu'il est impossible qu'une hypothèse unique convienne à tous les cas.

Exemple 2

Dans l'exemple précédent, la variable indépendante ainsi que les observations étaient rangées selon le même ordre, de sorte qu'on a supposé, pour l'établissement des modèles de régression, que les domaines de définition de la variable indépendante ne se chevauchaient pas. Bien que ce cas soit le plus courant, il arrive que les domaines de définition de la variable indépendante se chevauchent. Cela peut par exemple se produire dans l'étude de la stabilité dans le temps de la relation entre deux variables lorsque la valeur de la variable indépendante n'est pas contrôlée et que la personne qui effectue les mesures est remplacée à un certain moment par une autre personne.

Les données générées artificiellement par Quandt (1958) ont été analysées en supposant les options précisées dans le tableau 3. La figure 4 représente les deux graphiques élaborés respectivement pour décrire les données et les courbes de lissage d'une part et la fonction de vraisemblance relative d'autre part. Des symboles distincts ont été employés pour représenter les données et les courbes de lissage correspondant aux deux segments et la fonction de



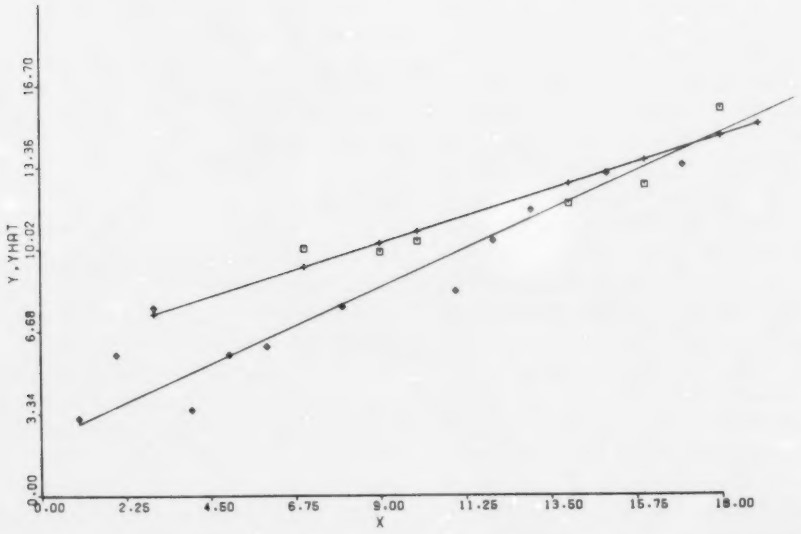


Figure 4. Représentation graphique des données avec les courbes de lissage et la fonction de vraisemblance relative, exemple 2.

vraisemblance relative est représentée de la même manière que dans le cas où un seul graphique est élaboré. Le modèle qui correspond à ce graphique est le modèle $M_e(1,1|12)$, avec $\hat{y} = 2.22 + 0.69x$ et $\hat{y} = 5.91 + 0.48x$ étant les équations des droites de lissage qui correspondent respectivement aux échantillons 1 à 12 et 13 à 20.

Exemple 3

La méthode que nous avons appliquée jusqu'à maintenant ne prévoit pas de test visant à démontrer l'existence d'un point de transition. Elle suppose à priori l'existence de ce point. Pour voir ce qui se passe lorsqu'on applique cette méthode à un ensemble de données ne comportant pas de point de transition, nous avons appliqué le programme à un échantillon de 100 variables aléatoires générées à partir d'une distribution normale N(0,1). Les options retenues sont celles qui sont précisées au tableau 3 et les graphiques sont représentés à la figure 5. Puisque la fonction de vraisemblance relative prend une valeur supérieure ou égale à 0.30 pour toutes les valeurs possibles du point de transition, on en conclut qu'il ne peut exister aucun point de cette nature.

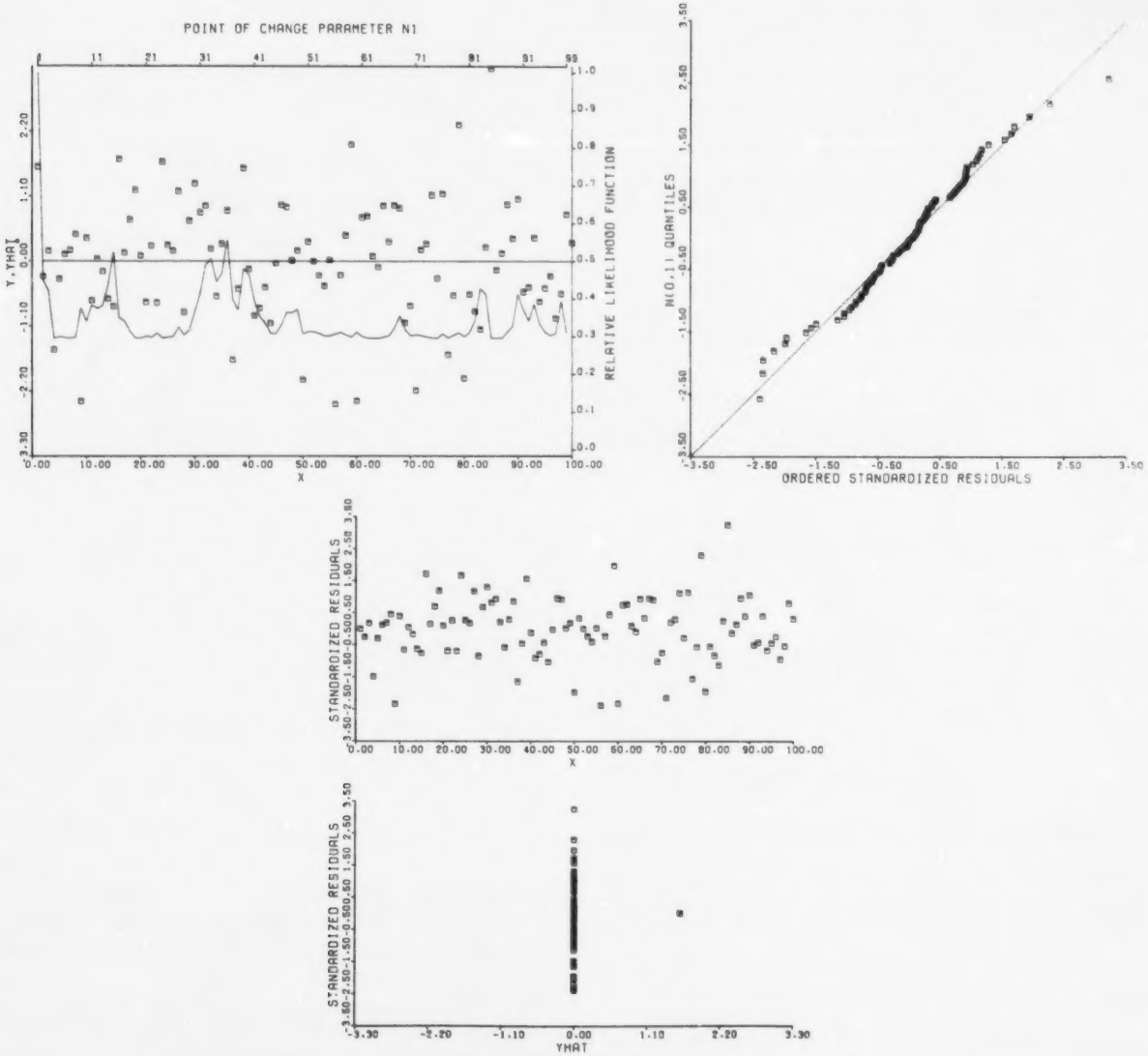


Figure 5. Représentation graphique des écarts normalisés, exemple 3.

Exemple 4

Bien que la méthode décrite ici ne s'applique en principe qu'au cas de l'existence d'un point de transition unique, on peut l'adapter de diverses manières au cas où il existe plusieurs points de transition. Une de ces adaptations est illustrée ici par l'exemple de la distribution verticale du pollen de type *Pinus* dans une carotte sédimentaire prélevée à Bog D Pond, Hubbard County, Minnesota. Les pourcentages ont été calculés à partir des valeurs indiquées par McAndrews (1966). Le pourcentage du pollen de type *Pinus banksiana/resinosa* décroît de haut en bas de la carotte alors que le pourcentage du pollen de type *Pinus strobus* se comporte de façon inverse. La somme de ces deux pourcentages devrait donc présenter deux points de transition. Le modèle définitif qui représente la somme de ces deux pourcentages a été établi après trois passages du programme. On a procédé de la façon ci-après :

1) À partir de la totalité des 49 échantillons, on a établi le modèle $M_e(2,1|44)$.

2) L'ensemble de données qui correspond aux échantillons 1 à 44 est décrit par le modèle $M_u(0,0|16)$.

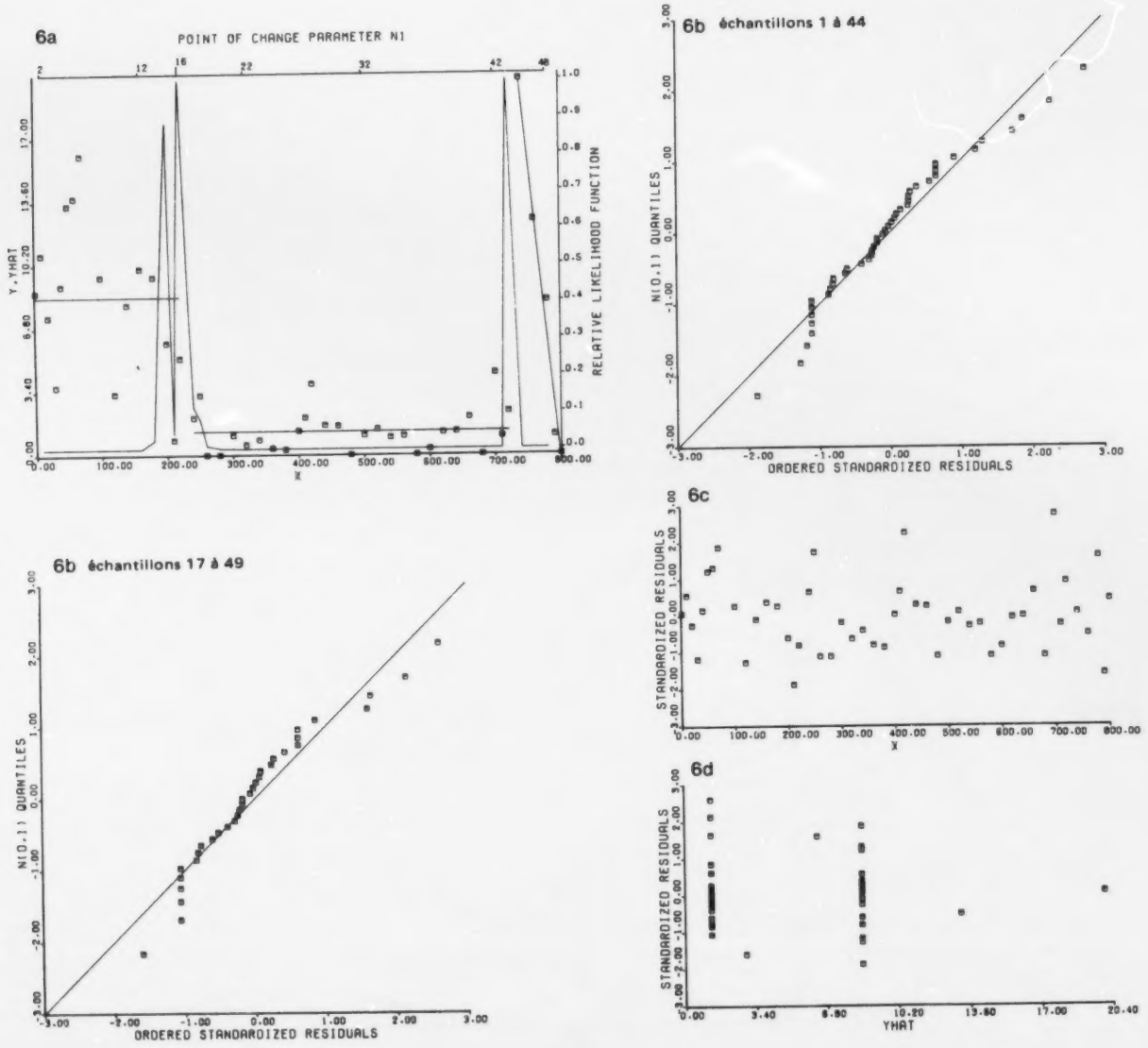


Figure 6. Graphiques tracés à partir des passages 2 et 3 du programme, exemple 4.

3) Pour vérifier que le point de transition inférieur correspond bien à l'échantillon 44, on a analysé une suite (à savoir les échantillons 17 à 49) comportant un seul point de transition. On a obtenu ainsi le modèle $M_u(0,1|44)$.

Les options retenues sont celles du tableau 3. Les graphiques produits lors des deux derniers passages du programme étant tous à la même échelle, on a pu découper et recoller les régions correspondant respectivement aux échantillons 1 à 16 et 17 à 49 (figures 6a,c,d). Le graphique Q-Q est représenté à la figure 6b. En choisissant les valeurs appropriées obtenues lors des deux derniers passages, le modèle définitif est défini par les relations ci-après

$$\hat{y}_i = 8.46 \qquad i = 1,2...16$$

$$= 1.25 \qquad i = 17,18...44$$

$$= 273 - 0.34\, x_i \qquad i = 45,46...49$$

avec $\hat{\sigma}_1^2$=16.52, $\hat{\sigma}_2^2$=1.27 et $\hat{\sigma}_3^2$=2.62. Le calcul de $100R^2$ pour trois segments se fait de la même façon que pour deux; cela donne la valeur $100R^2$ = 75.5. Bien que les écarts résiduels comportent encore un certain nombre de pointes, les variations importantes ont été repérées. En outre, le niveau de signification mesuré lors du test des suites appliqué au modèle à trois segments est égal à 0.20. Les valeurs nécessaires au calcul de $100R^2$ et du niveau de signification ont été obtenues lors de l'un des trois passages du programme.

Il n'est pas toujours possible d'obtenir, dans le cas où il existe plusieurs points de transition, une estimation satisfaisante de tous ces points. Cette situation peut notamment se rencontrer lorsque les variations sont du même ordre de grandeur. Toutefois, lorsqu'on parvient, par l'étude du diagramme de dispersion, à décomposer l'ensemble de données en segments contenant chacun un seul point de transition, on peut souvent obtenir une estimation satisfaisante grâce à une série de passages du programme.

ANALYSE

Les bornes supérieures des degrés des polynômes doivent être choisies avec soin pour éviter que la courbe de lissage tienne compte de faibles variations dans les données. Comme on l'a déjà mentionné, la courbe de lissage doit être étudiée en vue de déterminer les écarts éventuels par rapport aux hypothèses. Lorsqu'on attribue à la borne supérieure du degré du polynôme une valeur trop élevée dans l'un des deux segments, l'estimation du point de transition devient fortement dépendante de l'existence de points isolés dont la valeur diffère nettement de celle des points voisins. Cette caractéristique de la méthode est évidente lorsqu'on examine les graphiques.

Il est essentiel que la méthode permette de tenir compte de l'inégalité des variances dans les deux segments puisqu'il arrive souvent qu'une perturbation subie par un système entraîne une augmentation de la variance. La différence entre les variances dans les deux segments de la courbe de lissage n'apparaît pas dans les graphiques qui représentent les écarts résiduels puisque ces derniers sont normalisés. Les fluctuations de la variance qui ne sont pas dues simplement au fait que cette dernière prend une valeur constante différente dans chaque segment peuvent être détectées à partir de la représentation graphique des écarts résiduels et du test des suites.

Les exemples que nous avons présentés ne montrent pas comment le programme s'applique à plus d'une paire de variables. Le cas de l'exemple 4, qui a été traité en trois passages sans production de graphiques, aurait pu être traité en un seul passage avec production de graphiques. Chaque fois, ce sont les deux mêmes colonnes de données qui sont définies. La définition de paires de colonnes différentes donnerait bien à une situation plus habituelle. Ainsi, pour l'exemple 4, les colonnes de données auraient pu être constituées des trois paires : 1) profondeur et concentration du pollen de type *Pinus-banksiana/resinosa*, 2) profondeur et concentration du pollen de type *Pinus strobus* et 3) profondeur et somme des concentrations des pollens des types *Pinus-banksiana/résinosa* et *Pinus strobus*.

REMERCIEMENTS

L'algorithme a été codé et le programme testé par J. Hodson. Les sous-programmes de traçage de graphiques ont été rédigés à l'origine par D. Pateman; une version antérieure d'un programme semblable a été rédigée par M. Fellowes.

RÉFÉRENCES

Draper, N.R. et H. Smith. 1981. *Applied Regression Analysis*. 2e éd., John Wiley and Sons, Inc., New York.

Esterby, S.R. et A.H. El-Shaarawi. 1980. Examples of inferences about the point of change in a regression model. Non publié. Institut national de recherches sur les eaux, Burlington (Ontario).

Esterby, S.R. et A.H. El-Shaarawi. 1981a. Inference about the point of change in a regression model. J. Royal Stat. Soc. Series C, 30(3): 277-285.

Esterby, S.R. et A.H. El-Shaarawi. 1981b. Likelihood inference about the point of change in a regression regime. J. Hydrol. 53: 17-30.

McAndrews, J.H. 1966. Postglacial history of prairie, savanna and forest in Northwestern Minnesota. Mem. Torrey Bot. Club, Vol. 22, N° 2, 72 p.

Quandt, R.E. 1958. The estimation of parameters of a linear regression system obeying two separate regimes. J. Am. Stat. Assoc. 53: 873-880.

Robbins, J.A., D.N. Edgington et A.L.W. Kemp. 1978. Comparative ^{210}Pb, ^{137}Cs, and pollen geochronologies of sediments from Lakes Ontario and Erie. Quat. Res. 10(2): 256-278.

ANNEXE

La normalisation de la variable indépendante par soustraction de la moyenne et division par l'écart type, où la moyenne et l'écart type sont calculés à partir de toutes les valeurs de n, s'est révélée suffisante. Le processus de normalisation adopté, analogue à l'utilisation de la matrice de corrélation pour le calcul d'une fonction de régression unique, consiste à normaliser chaque segment indépendamment. Toutefois, cette méthode augmente notablement la complexité des calculs puisque les échantillons de chaque segment varient en fonction de n_1. On adopte donc, pour un segment particulier, la courbe de lissage

$$y_i = \beta_0 + \beta_1 \left(\frac{x_i - \bar{x}}{s}\right) + \ldots + \beta_d \left(\frac{x_i - \bar{x}}{s}\right)^d$$

et pour exprimer les valeurs de $\theta_0, \theta_1 \ldots \theta_d$ qui correspondent au modèle en fonction des variables d'origine, on calcul $\underline{\theta} = (\theta_0, \theta_1 \ldots \theta_d)'$ à partir de

$$\underline{\theta} = A\underline{\beta}^*$$

où $$\underline{\beta}^* = \begin{bmatrix} \beta_0 \\ \beta_1/s \\ \ldots \\ \beta_d/s^d \end{bmatrix}$$

A est une matrice (d+1) x (d+1) dont les éléments doivent être calculés. Ces éléments sont égaux aux termes du développement en série du binôme $(1-\bar{x})^{j-1}$ pour j = 1,2. . .d+1. Les j premiers éléments de la j^e colonne de A sont donnés par les termes de ce développement en série, en ordre inverse, et les derniers (d–j+1) éléments sont nuls. Ainsi, pour d=2, on a

$$\begin{bmatrix} \theta_0 \\ \theta_1 \\ \theta_2 \end{bmatrix} = \begin{bmatrix} 1 & -x & x^2 \\ 0 & 1 & -2x \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1/s \\ \beta_2/s^2 \end{bmatrix}$$

Le coefficient de corrélation multiple, R^2, est égal par définition à la somme des carrés attribuée à la régression divisée par la somme totale des carrés des écarts (TSS). Dans le modèle que nous étudions ici, le numérateur de R^2 n'est pas simplement égal à la somme des carrés attribuées à la régression calculées séparément pour chaque segment puisque cette somme ne tient pas compte de toutes les composantes du modèle. Pour le voir, il suffit de décomposer la somme totale des carrés des écarts comme suit :

$$TSS = \sum_{i=1}^{n} (y_i - \bar{y})^2 = \sum_{i=1}^{n_1} (y_i - \bar{y})^2 + \sum_{i=n_1+1}^{n} (y_i - \bar{y})^2$$

$$= TSS_1 + TSS_2$$

Mais

$$TSS_1 = \sum_{i=1}^{n_1} (y_i - \bar{y}_1)^2 + \sum_{i=1}^{n_1} (\bar{y}_1 - \bar{y})^2$$

et le premier terme est égal à

$$\sum_{i=1}^{n_1} (y_i - \hat{y}_{1i})^2 + \sum_{i=1}^{n_1} (\hat{y}_{1i} - \bar{y}_1)^2$$

où l'indice 1 désigne une valeur calculée à partir des échantillons du segment 1. Les deux termes ci-dessus correspondent respectivement à la somme des carrés des écarts et à la somme des carrés attribuée à la régression pour le segment 1. En appliquant la même décomposition au segment 2, puis en recombinant et en groupant les termes correspondants des deux segments, on obtient

$$TSS = \{ResSS_1 + ResSS_2\} + \{RegSS_1 + RegSS^2\} + M$$

où $M = n_1(\bar{y}_1 - \bar{y})^2 + (n-n_1)(\bar{y}_2 - \bar{y})^2$ et ResSS et RegSS désignent respectivement la somme des carrés des écarts et la somme des carrés attribuée à la régression. On a donc

$$100R^2 = 100\{TSS - (ResSS_1 + ResSS_2)\}/TSS$$